Linear Convergence of Decentralized FedAvg for PL Objectives: The Interpolation Regime

Shruti Maralappanavar *mshruti32@gmail.com*
Department of Electrical, Electronics and Communication
IIT Dharwad
Dharwad, India

Prashant Khanduri *khanduri.prashant@wayne.edu*
Department of Computer Science
Wayne State University
Detroit, MI, USA

B. N. Bharath *bharathbn@iitdh.ac.in*
Department of Electrical, Electronics and Communication
IIT Dharwad
Dharwad, India

Reviewed on OpenReview: *https://openreview.net/pdf?id=Og3VxBFhwj*

Abstract

Federated Learning (FL) is a distributed learning paradigm where multiple clients each having access to a local dataset collaborate to solve a joint problem. Federated Averaging (FedAvg) the algorithm of choice has been widely explored in the classical *server* setting where the server coordinates the information sharing among clients. However, the performance of FedAvg in the *decentralized* setting where only the neighboring clients communicate with each other depending on the network topology is not well understood, especially in the interpolation regime, a common phenomenon observed in modern overparameterized neural networks. In this work, we address this challenge and perform a thorough theoretical performance analysis of FedAvg in the interpolation regime under *decentralized* setting. We consider a class of non-convex functions satisfying the Polyak-Łojasiewicz (PL) inequality, a condition satisfied by overparameterized neural networks. For the first time, we establish that *Decentralized* FedAvg achieves linear convergence rates of $\mathcal{O}(T^2 \log(1/\epsilon))$, where ϵ is the solution accuracy, and T is the number of local updates at each client. We also extend our analysis to the classical *Server* FedAvg and establish a convergence rate of $\mathcal{O}(\log(1/\epsilon))$ which significantly improves upon the best-known rates for the simpler strongly-convex setting. In contrast to the standard FedAvg analyses, our work does not require bounded heterogeneity and gradient assumptions. Instead, we show that sample-wise (and local) smoothness of the local objectives suffice to capture the effect of heterogeneity. Experiments on multiple real datasets corroborate our theoretical findings.

1 Introduction

In the age of Bigdata, Federated Learning (FL) provides machine learning (ML) practitioners with an indispensable tool for solving large-scale learning problems. FL is a distributed machine learning scenario that allows the edge devices to learn a shared model while maintaining the training data at the edge devices (Konečnỳ et al., 2016; McMahan et al., 2017). This avoids the need to share the data with a central server and preserves the privacy of individual clients (edge devices). Assuming a supervised learning setting, where each of the N distinct clients have access to some local data $(\boldsymbol{x}, y) \sim \mathcal{D}_k$ from distribution $\mathcal{D}_k$ for $k \in [N]$

aim to solve:

$$\texttt{FL Problem:} \min_{\boldsymbol{w}\in\mathbb{R}^d} \Phi(\boldsymbol{w}) := \frac{1}{N}\sum_{k=1}^{N} \Phi_k(\boldsymbol{w}),$$

where $\Phi_k(\boldsymbol{w}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_k} l_k(f_{\boldsymbol{w}}(\boldsymbol{x}), y)$ is the expected loss at client $k \in [N]$ for the input feature $\boldsymbol{x} \in \mathcal{X}$, and the corresponding label $y \in \mathcal{Y}$. Here, $f_{\boldsymbol{w}}(\boldsymbol{x})$ is the model's output with parameters $\boldsymbol{w} \in \mathbb{R}^d$. In this work, we focus on solving the `FL Problem` in the interpolation regime (cf. Assumption 1), an assumption usually satisfied by overparameterized models (Bassily et al., 2018).

The de-facto standard for solving the above `FL Problem` is the simple Federated Averaging (FedAvg) algorithm (McMahan et al., 2017). In recent years, many works have attempted to characterize the convergence of FedAvg under different settings (Stich, 2018; Li et al., 2019a; Woodworth et al., 2020a; Ma et al., 2018; Yu et al., 2019b). For example, the authors in Stich (2018) show a convergence rate of $\mathcal{O}\left(1/N\epsilon\right)$ for minimizing strongly convex functions while Haddadpour & Mahdavi (2019) establishes similar rates for minimizing functions satisfying Polyak-Lojasiewicz (PL) condition. For minimizing non-convex smooth functions, FedAvg achieves a convergence rate of $\mathcal{O}(1/N\epsilon^2)$ (Karimireddy et al., 2020; Woodworth et al., 2020b), where, ϵ refers to the desired solution accuracy.

Note that the majority of studies have focused on FedAvg in the classical server-based setting (referred to as *Server* FedAvg), where a central server aggregates information. In this approach, clients compute model updates and send them to the server, potentially leading to communication bottlenecks and delays. Moreover, an attack on the server could compromise the privacy of the aggregated model. Additionally, in many practical learning scenarios, access to a central server may not be feasible. For such cases, the alternative is *Decentralized* FedAvg. In this decentralized setting, global aggregation is replaced by local aggregation, where each client updates its model based on connections with neighboring clients. While the Decentralized FedAvg algorithm has been examined in earlier works (Koloskova et al., 2019; Li et al., 2019b), its linear (fast) convergence has only been established for the strongly convex setting in the interpolation regime (cf. Assumption 1). In practice, it has been observed that even for non-convex settings *Decentralized* FedAvg converges at a much faster rate compared to the rates demonstrated in these works. To illustrate this fact, in Fig. 1 we plot the behavior of the training loss (on a log scale) as a function of communication rounds for *Decentralized* FedAvg to solve a classification task on MNIST data set utilizing overparamterized models, i.e., in the interpolation regime (for experimental details please see Section 5). It is clear from the plot that in the interpolation regime the loss decreases linearly as a function of communication rounds. This suggests that the standard analyses of *Decentralized* FedAvg lack a theoretical foundation to explain its linear convergence.

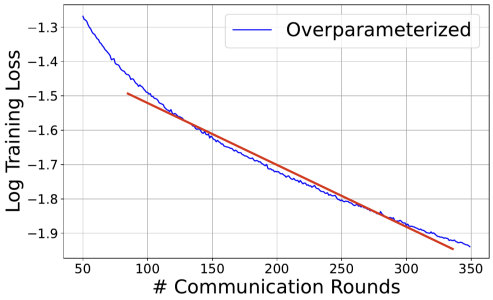

Figure 1: log-training loss vs communication rounds for overparameterized Deep Neural Networks (DNNs) and a simple regression model.

In this work, we attempt to fill these gaps and perform a thorough theoretical analysis of *Decentralized* FedAvg in the interpolation regime where the local nodes communicate over an undirected graph. Under this setting, we establish the linear convergence of FedAvg for minimizing a class of non-convex functions satisfying the PL inequality. We also extend our analysis to the classical *Server* FedAvg and establish improved guarantees compared to the state-of-the-art. We note that PL inequality plays a key role in the training of overparameterized systems. Specifically, many works have shown that the loss function of an overparameterized neural network satisfies the PL inequality (Bassily et al., 2018; Liu et al., 2020). Furthermore, our analysis reveals that the standard but restrictive assumptions of bounded gradients (Yu et al., 2019b; Stich, 2018; Li et al., 2019a; Koloskova et al., 2020), heterogeneity (Yu et al., 2019a; Woodworth et al., 2020b; Yu et al., 2019a; Wang et al., 2021; Sery et al., 2021), and variance (Woodworth et al., 2020b; Qu et al., 2020) can be avoided while guaranteeing this linear convergence of *Decentralized* FedAvg.

Table 1: Comparison with existing works. Here, SC, C and NC represent Strongly convex, Convex and Non-convex settings, respectively.

ALGORITHM	CONVERGENCE	EXTRA ASSUMPTIONS	SETTING
Local SGD (Stich, 2018) (s)	$\mathcal{O}(1/N\epsilon)$	Bounded gradient	SC
Local SGD (Yu et al., 2019b) (s)	$\mathcal{O}(1/N\epsilon^2)$	Bounded variance, smoothness	NC
Local SGD (Haddadpour et al., 2019) (s)	$\mathcal{O}(1/N\epsilon)$	Bounded variance, smoothness	NC
FedAvg (Qu et al., 2020) (s)	$\mathcal{O}(T\log(1/\epsilon))$	Bounded Gradient, Bounded Variance	Overparameterized SC
Local SGD (Woodworth et al., 2020b) (s)	$\mathcal{O}(1/N\epsilon^2)$	Bounded Variance	C
Local SGD (Woodworth et al., 2020a) (s)	$\mathcal{O}(1/N\epsilon^2)$	Bounded Variance	C
PR-SGD (Yu et al., 2019a) (s)	$\mathcal{O}(1/N\epsilon^2)$	Bounded Variance	NC
FedAvg (Karimireddy et al., 2020) (s)	$\mathcal{O}(1/N\epsilon^2)$	Bounded gradient dissimilarity Bounded heterogeneity	NC
SGD (Koloskova et al., 2020) (s)	$\mathcal{O}(T\log(T/\epsilon))$	Smoothness	Overparameterized SC
OUR WORK (s)	$\mathcal{O}(\log(1/\epsilon))$	Smoothness	Overparameterized NC
NFSGD (Haddadpour & Mahdavi, 2019)	$\mathcal{O}(1/N\epsilon^2)$	Bounded local variance	NC
DECENTRALIZED SGD (Koloskova et al., 2020)	$\mathcal{O}(T\log(T/\epsilon))$	Smoothness	Overparameterized SC
DECENTRALIZED GD(Song et al.)	$\mathcal{O}(\log(1/\epsilon))$	Smoothness	Overparameterized NC
OUR WORK	$\mathcal{O}(T^2\log(1/\epsilon))$	Smoothness	Overparameterized NC

Contributions. Here, we list the major contributions of our work:

1. First, we consider the *decentralized* setting where N distributed clients communicate over an undirected graph. We show that to achieve an ϵ-accurate solution *Decentralized* FedAvg requires $R \sim \mathcal{O}\left(T^2 \log(1/\epsilon)\right)$ rounds of communication. We also characterize the effect of the network topology on the performance of *Decentralized* FedAvg.

2. For the classical *Server* FedAvg also we establish linear convergence rates of $R \sim \mathcal{O}(\log(1/\epsilon))$. Compared to the best-known result in Koloskova et al. (2020) where a simpler strongly-convex problem is considered (see section C for details), we get improved rates independent of the number of local updates T (cf. Table 1 and Section 1.1 for a comparison).

3. Our theoretical results do not require assumptions on the boundedness of heterogeneity, gradients, and stochastic variance. We show that sample-wise smoothness of the stochastic loss functions suffices to capture the effect of data heterogeneity among different clients while avoiding the need to impose the restrictive bounded gradient and variance assumptions.

4. Finally, we present experimental results on various data sets such as CIFAR-10, Shakespeare, MNIST, and FMNIST, and corroborate our theoretical findings under the *decentralized* settings. Through our experiments, we show that an optimal number of local updates T_{th} exists and increasing T beyond T_{th} hurts the performance of FedAvg in both *server* and *decentralized* settings.

1.1 Related Work

After the introduction of the FedAvg (McMahan et al., 2017), multiple works have analyzed the convergence of FedAvg in the server setting with homogeneous data, i.e., when the data is i.i.d across clients (see (Stich, 2018; Wang & Joshi, 2018; Khaled et al., 2019; Yu et al., 2019b; Wang et al., 2019; Yang et al., 2021)). The authors in (Stich, 2018) were the first to obtain a rate of $\mathcal{O}(1/N\epsilon)$ for strongly convex and smooth problems for achieving an ϵ-accurate solution (cf. Definition 2). Later (Haddadpour et al., 2019; Haddadpour & Mahdavi, 2019) proved a similar result but for non-convex functions satisfying PL inequality. The analysis of FedAvg for the general non-convex settings was first performed in (Yu et al., 2019b) where the authors establish a rate of $\mathcal{O}(1/N\epsilon^2)$ to reach an ϵ-stationary point where the stationarity is measured with respect to the gradient norm of the loss funciton.

There are a few works that have analyzed the performance of Fedvg in *decentralized* settings as well. One of the initial works, (Lian et al., 2017) considered a decentralized parallel SGD (D-PSGD) and provided a convergence rate of $\mathcal{O}(1/N\epsilon^2)$ to an ϵ-stationary point to minimize smooth non-convex functions. Later,

(Haddadpour & Mahdavi, 2019) analyzed the convergence of FedAvg under both *server* and *decentralized* settings with bounded gradient dissimilarity assumption. The authors also established a convergence rate of $\mathcal{O}(1/N\epsilon^2)$ to an ϵ-stationary point to minimize nonconvex functions in both the *server* and *decentralized* settings. The authors in (Yu et al., 2019a) also extended the analysis of Momentum SGD to decentralized networks and established convergence of $\mathcal{O}(1/N\epsilon^2)$ to an ϵ-stationary point to minimize non-convex functions. Recently, the authors in (Song et al.) established a linear convergence rate of $\mathcal{O}(\log(1/\epsilon))$ to an ϵ-accurate solution for a decentralized gradient descent algorithm in the overparameterized regime. But, it is well known that in the deterministic setting distributed algorithms for minimizing PL loss functions are capable of achieving linear convergence (even in the non-interpolated regime). Moreover, for data-centric problems, deterministic algorithms involve computations of large batch gradients during each update, making them impractical for large-scale problems. Further, the local nodes conduct only a single local update. However, in this work, we consider a *stochastic problem* where each agent conducts *multiple local updates* within each communication round. Moreover, it is an open problem to establish if (stochastic) *Decentralized* FedAvg can guarantee linear convergence in the interpolation regime for non-strongly convex losses. All the above works provide a sublinear rate of convergence for *Decentralized* FedAvg, however, as illustrated in Fig. 1, *Decentralized* FedAvg converges at a much faster rate in practice. To understand this behavior of *Decentralized* FedAvg, in this work, we analyze the performance of *Decentralized* FedAvg for minimizing a special class of non-convex functions satisfying PL inequality under the interpolation regime. We note that overparameterized neural networks/systems usually operate in the interpolation regime while their loss functions have been shown to satisfy the PL inequality.

Similarly, the authors in (Koloskova et al., 2020) have also established the linear convergence of FedAvg in the *decentralized* setting for minimizing strongly-convex losses in an overparameterized setting. The above works only focus on the analysis of FedAvg for the strongly-convex objectives in the overparameterized regime while we focus on the more general class of non-convex functions satisfying the PL inequality. Note that for strongly-convex objectives the local clients will all share a unique minimum, this implies that even if there is no communication, FedAvg will converge with the same rate to the local optimal. Importantly, our analysis improves the analyses of (Qu et al., 2020) and (Koloskova et al., 2020), and establishes better dependence on the local updates on the performance of FedAvg (Please see Table 1). Moreover, compared to other works that assume restrictive bounded gradient, heterogeneity, and variance assumptions, we show that such assumptions can be avoided by using a sample-wise smoothness assumption. The authors in (Qu et al., 2020) show linear speedup for strongly convex and convex functions. Table 1 presents a summary of the above discussion.

The *Decentralized* SGD algorithm is extensively studied under various conditions which consists of time-varying network graphs (Koloskova et al., 2020), momentum updates (Lin et al., 2021), asynchronous model updating (Nadiradze et al., 2021). To handle the heterogeneity in data across the clients, various tracking algorithms such as gradient tracking, model tracking (Yue Liu & Stich, 2024; Aketi et al., 2024) and momentum tracking (Takezawa et al., 2023) have also been proposed. In separate line of works, the authors in (Zhu et al., 2023b;a) prove generalization guarantees for *Decentralized* SGD algorithm. The authors in (Nadiradze et al., 2021) along with asynchronous updates consider quantization and multiple local steps. Some recent works such as (Beznosikov et al., 2022) have looked at the stochastic extragradient method with time varying networks for the decentralized methods.

Notations: We use bold small letters to denote vectors and capital bold letters for matrices. We denote the expected value of a random variable X by $\mathbb{E}[X]$. We denote l_2-norm by $\|\cdot\|_2$, Frobenius norm by $\|\cdot\|_F$ and operator norm by $\|\cdot\|_{op}$. $\langle\cdot,\cdot\rangle$ denotes the inner product. The cardinality of any set $\mathcal{B}$ is represented by $|\mathcal{B}|$. We use the standard notation $\mathcal{O}(n)$ to denote the order of n. For a vector-valued function $\Phi(\boldsymbol{w})$, the gradient is denoted by $\nabla\Phi(\boldsymbol{w}) \in \mathbb{R}^d$, and the Hessian is denoted by $\nabla^2\Phi(\boldsymbol{w}) \in \mathbb{R}^{d\times d}$. We use $\mathbf{1}$ to represent a column vector with all ones. We use $[N]$ to denote the set $\{1,\ldots,N\}$.

2 The *Decentralized* FedAvg Algorithm

In many practical settings, the central server is absent, and the clients are required to communicate and update the model weights in a decentralized manner while communicating with only the neighboring nodes. In this work, we consider such a setting and study *Decentralized* FedAvg, an extension of FedAvg to the decentralized setting. The *decentralized* setting consists of N distributed edge devices which are represented using a connectivity graph $\mathcal{G} \in (\mathcal{V}, \mathcal{E})$. Here, $\mathcal{V} \in [N]$ is the vertex set or clients, and $\mathcal{E} \subseteq \{\mathcal{V} \times \mathcal{V}\}$ represents the edges of the graph. Any edge $(i, j) \in \mathcal{E}$ represents a connection between node i and j. Further, the connections are represented using mixing matrix $P = [p_{i,k}] \in \mathbb{R}^{N \times N}$, where $p_{i,k} = 0$ if there is no edge between node i and k i.e., $(i, k) \notin \mathcal{E}$, else $p_{k,i} > 0$. Unlike many existing works on *decentralized* settings (Lian et al., 2017), here we consider a general framework where each client performs T rounds of local updates. *Decentralized* FedAvg is presented in **Algorithm** 1 while in the following, we provide an outline:

1. `Initialization:` Each client $k \in [N]$ initializes the model parameters denoted by $\underline{\boldsymbol{w}}_k^0$. See Step 1 of Algorithm 1.

2. `Local updates:` Each client performs T steps of SGD starting from the model parameters obtained by the aggregation of the updates from neighbouring clients. Towards computing the stochastic gradient, each client $k \in [N]$ uniformly randomly samples a batch of data of size b denoted by $\mathcal{B}_k^{r,t}$, and then computes the gradient. The resulting model parameters after T local rounds in the r-th global round are denoted by $\boldsymbol{w}_k^{r,T}$ which is sent to all the neighbouring clients of k. See Steps 6 and 7 of Algorithm 1. Note that this procedure is similar to the local update step in the FedAvg case.

3. `Aggregation:` In the r-th global communication round, each client $k \in [N]$ computes a local average of the model parameters received by its neighbors. The aggregate model is denoted by $\underline{\boldsymbol{w}}_k^r$. The steps (2) and (3) above are repeated for R rounds. See Steps 11, 12 of Algorithm 1.

2.1 Assumptions

In this subsection, we present the assumptions and definitions used in the analysis of the *Decentralized* FedAvg algorithm.

Definition 1. *(L-Smoothness): The function $\Phi(\boldsymbol{u})$ is said to be L smooth if there exists a constant $L > 0$ such that $\|\nabla\Phi(\boldsymbol{u}_1) - \nabla\Phi(\boldsymbol{u}_2)\|_2 \leq L\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_2$ for any $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^d$. Note that this further implies that $\Phi(\boldsymbol{u}_1) \leq \Phi(\boldsymbol{u}_2) + \langle \nabla\Phi(\boldsymbol{u}_2), \boldsymbol{u}_1 - \boldsymbol{u}_2 \rangle + \frac{L}{2}\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|^2$ for any $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^d$.*

Definition 2. *(ϵ-accurate solution): A stochastic algorithm is said to achieve an ϵ-accurate solution in r rounds if $\mathbb{E}[\Phi(\boldsymbol{w}^r) - \Phi(\boldsymbol{w}^*)] \leq \epsilon$, where the expectation is taken over the stochasticity of the algorithm and $\boldsymbol{w}^* \in \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \Phi(\boldsymbol{w})$.*

Assumption 1. *(Interpolation): We say that the model parameters is operating in the interpolation regime if there exists a $\boldsymbol{w} \in \mathbb{R}^d$ such that the per sample loss $\Phi_{k,j}(\boldsymbol{w}) = 0$ for all samples $j \in [b]$.*

Assumption 2. *(PL inequality): The joint objective $\Phi(\boldsymbol{v})$ satisfies the PL inequality, i.e., $\|\nabla\Phi(\boldsymbol{v})\|^2 \geq \mu\Phi(\boldsymbol{v})$ for some $\mu > 0$ and for all $\boldsymbol{v} \in \mathbb{R}^d$. Further, the local loss functions $\Phi_k(\boldsymbol{v})$ for all $k = 1, 2, \ldots, N$ are also assumed to satisfy the PL inequality, henceforth referred to as local PL inequality, i.e., $\|\nabla\Phi_k(\boldsymbol{v})\|^2 \geq \mu_k\Phi_k(\boldsymbol{v})$ for some $\mu_k > 0$ and for all $\boldsymbol{v} \in \mathbb{R}^d$.*

Assumption 3. *(Sample-wise, Local and Global smoothness): The functions $\Phi_{k,j}(\cdot)$ for all $j \in [b], k \in [N]$ are assumed to be $l_{k,j}$-smooth. The local functions $\Phi_k(\cdot)$ for all $k \in [N]$ are assumed to be L_k-smooth. The above assumptions imply $\|\nabla\Phi_{k,j}(\boldsymbol{v})\|^2 \leq 2l_{k,j}\Phi_{k,j}(\boldsymbol{v})$ and $\|\nabla\Phi_k(\boldsymbol{v})\|^2 \leq 2L_k\Phi_k(\boldsymbol{v})$ for all $k \in [N]$ and $j \in [b]$. We also assume the global loss $\Phi(\cdot)$ to be L-smooth.*

Assumption 4. *(Unbiasedness): We assume that the stochastic samples of the gradient and the loss function at each client $k \in [N]$ are unbiased, i.e., $\mathbb{E}[\nabla\Phi_{k,j}(\boldsymbol{w})] = \nabla\Phi_k(\boldsymbol{w})$ and $\mathbb{E}[\Phi_{k,j}(\boldsymbol{w})] = \Phi_k(\boldsymbol{w})$ for any $j \in [b]$ and $\boldsymbol{w} \in \mathbb{R}^d$.*

Most of the above assumptions, including interpolation, PL inequality, and smoothness are standard assumptions made in various works in the past (Bassily et al., 2018; Karimi et al., 2016; Ma et al., 2018;

Algorithm 1 *Decentralized* FedAvg

Initialize $\{\boldsymbol{w}_k^{0,0} = \underline{\boldsymbol{w}}_k^0\}$, $\boldsymbol{w}_k \in \mathbb{R}^d$ for $k \in [N]$
for $r = 0, 1, \ldots, R-1$ **do**
  Initialize $\underline{\boldsymbol{w}}_k^r$ at device $k \in [N]$
  for $t = 0, 1, \ldots, T-1$ **do**
    for devices $k \in [N]$ **do**
      Sample a batch $\mathcal{B}_k^{r,t}$ and $|\mathcal{B}_k^{r,t}| = b$
      SGD step on $\boldsymbol{w}_k^{r,t}$ for $k \in [N]$: $\boldsymbol{w}_k^{r,t+1} = \boldsymbol{w}_k^{r,t} - \frac{\eta}{b} \sum_{j \in \mathcal{B}_k^{r,t}} \nabla \Phi_{k,j} \left(\boldsymbol{w}_k^{r,t} \right)$
    end for
  end for
  Receive $\boldsymbol{w}_k^{r,T}$ from clients $k \in [N]$
  Aggregation step : $\underline{\boldsymbol{w}}_k^{r+1} = \sum_{i \in \mathcal{N}_k} p_{k,i} \boldsymbol{w}_i^{r,T}$, for $k \in [N]$
end for

Nguyen & Mondelli, 2020). For example, the authors in (Bassily et al., 2018; Liu et al., 2022a; Karimi et al., 2016; Haddadpour et al., 2019) assume PL inequality along with sample-wise smoothness to prove linear convergence of FedAvg in the interpolation regime. It is also important to note that the overparameterized systems satisfy PL inequality (Liu et al., 2020; Nguyen & Mondelli, 2020; Nguyen et al., 2021; Allen-Zhu et al., 2019; Liu et al., 2022a), and hence plays a crucial role in the analysis of overparameterized systems (Liu et al., 2022a). Moreover, we note that the assumption of sample-wise smoothness is not very stringent since any neural network with a smooth activation function satisfies this assumption.

3 Convergence of *Decentralized* FedAvg

In this section, we prove that the *Decentralized* FedAvg algorithm converges linearly to the global optimum for any smooth non-convex function satisfying PL inequality in the interpolation regime. Compared to the classical *Server* FedAvg analyses this problem poses several challenges. In particular, unlike *Server* FedAvg in the *decentralized* setting each client has access to only parameters from its neighbours. This implies that for *Decentralized* FedAvg we need to handle two drift terms, namely *local drift* and the *global drift*. Local drift refers to the update at each client drifting away from the average obtained from the neighboring clients while the global drift refers to the average obtained from the neighboring clients drifting away from the global average. These two equations are coupled, and hence we use the Lyapunov based approach to show that both drift as well as the loss go down linearly. In addition to the Assumptions 1-4, our analysis also relies on the following assumption on the mixing matrix (Koloskova et al., 2020).

Assumption 5. *P is symmetric, i.e., $P = P^T$, and doubly stochastic, i.e., $P\mathbf{1} = \mathbf{1}$, $\mathbf{1}^T P = \mathbf{1}^T$.*

The above assumption covers all networks that are symmetric, for example, fully connected graph, ring graph, (Hua et al., 2022), etc. In the following, we provide the main result for the *Decentralized* FedAvg. The details of the proof are presented in Sec. B of the Appendix.

Theorem 1 (**Convergence of Algorithm** 2)**.** *Under Assumptions 1-5, after T local iterations, choosing $\eta = \mathcal{O}(1/T^2)$, we get*

$$\mathbb{E}[\Phi\left(\underline{\mathbf{w}}^{r+1}\right) + \theta \mathcal{D}_{r+1,0}] \leq \left(1 - \frac{\eta\mu}{16}\right) \mathbb{E}[\Phi\left(\underline{\mathbf{w}}^{r}\right) + \theta \mathcal{D}_{r,0}].$$

Proof: See Sec. B.2 in Appendix and Sec. 3.1 for a proof sketch. □

Theorem 1 establishes linear convergence of FedAvg in the *decentralized* setting. A somewhat related work is Koloskova et al. (2020), where they consider a class of μ-strongly convex functions in the decentralized setting, and showed that a linear convergence rate of $\mathcal{O}(\log(\frac{1}{\epsilon}))$ can be achieved by the decentralized FedAvg algorithm in the interpolation regime. Note that in a strongly convex setting, all clients share the unique minima $\mathbf{x}^* = 0$ due to the interpolation assumption. In this case, clients do not need to communicate since

each client can run multiple local rounds to reach the global minima that minimize the average, and hence making decentralized or collaborative learning vacuous! On the other hand, for the PL setting, multiple local rounds may lead to different optimal points, making the problem more challenging. In order to prove linear convergence, the authors in Koloskova et al. (2020) consider bounding a part of the global drift term. More specifically, they consider $\left\|\bar{\mathbf{x}}^{r+1} - \mathbf{x}^*\right\|$ (Lemma 8), where $\bar{\mathbf{x}}^{r+1}$ is the average of the local updates and $\mathbf{x}^*$ is the global optimal. In our setting, we cannot consider $\mathbf{x}^*$ as there is no unique local or global optimal point. Furthermore, they use the property of strong convexity to bound $\left\|\bar{\mathbf{x}}^{r+1} - \mathbf{x}^*\right\|$; this cannot be done in our setting. Hence, our analysis is quite different from the existing work. See Sec. C for more details. Next, we characterize the sample complexity of *Decentralized* FedAvg.

Corollary 1. *Under Assumptions 1-5, to achieve an ϵ-accurate solution, Algorithm 1 requires $R = \mathcal{O}\left(T^2/\mu \left[\log\left(\mathbb{E}[\Phi\left(\underline{\boldsymbol{w}}^0\right)]/\epsilon\right)\right]\right)$ number of communication rounds.*

Proof: It is clear from equation 48 of Theorem 1 that η scales as $\frac{\mu}{\zeta_8 T^2}$. Thus, from a scaling point-of-view, using $\eta = \frac{\mu}{\zeta_8 T^2}$ in Theorem 1 and the fact that $(1-x) \leq e^{-x}$, we get

$$\mathbb{E}[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right) + \theta \mathcal{D}_{r+1,0}] \leq \exp\left(-\frac{R\mu}{\zeta_8 T^2}\right) \mathbb{E}\left(\Phi\left(\underline{\boldsymbol{w}}^0\right)\right).$$

From above we see that to obtain ϵ accuracy, we want $\exp\left(-\frac{R\mu}{\zeta_8 T^2}\right) \mathbb{E}\left(\Phi\left(\underline{\boldsymbol{w}}^0\right)\right) \leq \epsilon$. Now rearranging the above, and using the fact that $\mathbb{E}[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right) + \theta \mathcal{D}_{r+1,0}] \geq \mathbb{E}[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)]$ gives us the result in the corollary. □

Corollary 1 shows that even in the *decentralized* setting, FedAvg can achieve linear convergence. More importantly, the sum of the drift and the loss goes to zero linearly with R, as opposed to most of the existing work (Sun et al., 2021). Observe from Theorem 1 and the corollary above that an ϵ-accurate solution can be achieved if the number of global communications rounds R scales as $\mathcal{O}(T^2)$. For the strongly convex setting since the clients share a unique minima the impact of the local rounds T on the convergence performance is less severe as shown in (Koloskova et al., 2020). Since the local clients share a unique minimum, which allows the algorithm to converge at the same rate to the local optimal irrespective of the local updates and communication protocol used. We believe that the slightly worse convergence of our analysis is due to the non-convex functions satisfying the PL inequality. Importantly, as a special case of our analysis, for the classical FedAvg algorithm (with a server), we establish a convergence rate of $\mathcal{O}(\log(1/\epsilon))$ which is significantly better than that existing results (see Section 4). Note that one can optimize the number of local rounds that lead to faster convergence. However, this optimization is cumbersome in the *decentralized* setting, and hence the convergence depends on T.

Effect of Network Topology. The effect of *decentralized* clients is captured through the term involving λ_2. In order to explain the dependency of λ_2 on convergence, consider the case of $T = 1$, i.e., FedSGD. In this case, if $\lambda_2 \neq 0$ but closer to 1, then the learning rate is dominated by the term $(1-\lambda_2)/\text{constant}$. Thus, the learning rate is small in the *decentralized* setting (as opposed to $\lambda_2 = 0$). As a consequence, $(1 - \eta\mu/8)$ is closer to 1 leading to a slower convergence. In the extreme case of $\lambda_2 = 1$, i.e., fully disconnected graph, $\eta = 0$, which leads to divergence, as expected. Later, we perform experiments to show the effect of network topology on the convergence for different network settings.

3.1 Proof Sketch of Theorem 1

In contrast to the strongly convex setting of (Koloskova et al., 2020), our setting results in each client sharing different optimal points. As a consequence, the execution of multiple local updates within each communication round leads to the following drifts: (i) local drift: multiple local updates (see line 4 to line 8 of Algorithm 1) differ from the average obtained by the neighboring clients, and (ii) global drift: the local average obtained by using updates from the neighboring clients differ from the global average, i.e., the average obtained from all the clients. This implies that we need to control both global and local drifts. We handle this challenge by bounding the loss in terms of the drift term that captures both local and global drifts as mentioned in the Lemma below. We start by proving an upper bound on the average loss $\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right]$ in terms of the loss $\Phi\left(\underline{\boldsymbol{w}}^r\right)$ in the r-th communication round, and the drift $\mathcal{D}_{r,0}$, as shown in

the following Lemma.

Lemma 1. *The average loss is bounded in terms of the drift as follows*

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \leq \left(1-\frac{\eta\mu}{8}\right)\Phi\left(\underline{\boldsymbol{w}}^{r}\right)+\frac{6\eta L^2 T}{N}\mathcal{D}_{r,0}, \tag{1}$$

where the drift $\mathcal{D}_{r,0} := \sum_{k=1}^{N}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^{r,0}-\underline{\boldsymbol{w}}^{r,0}\right\|^2$*, and* η *is chosen according to equation 48.*

Proof: See Sec. B.2. □

It is easy to see from Lemma 1 that we can obtain the convergence result established in Theorem 1 provided the drift term on the right hand side of equation 1 is bounded in terms of loss. Towards this, first, we bound the drift term which depends on the average loss, leading to two coupled equations (see equation 1 and equation 2). We construct a single equation that is a linear combination of the two coupled equations and show that the linear combination goes to zero exponentially, leading to linear convergence of both drift as well as the loss function. In the following lemma, we provide a recursion of the drift in terms of the average loss and the past drift.

Lemma 2. *The drift is bounded in terms of* $\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)$ *as follows*

$$\mathcal{D}_{r,0} \quad \leq \quad \left(\left(1+\frac{1}{\psi}\right)\lambda_2^2+\eta^2\lambda_2^4\beta T^2 L_{max}^2\right)\mathcal{D}_{r-1,0}+2\eta^2\lambda_2^4\beta T^2 L_{max} N\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right], \tag{2}$$

where $\beta := \frac{4l_{max}(1+\psi)}{\mu_{min}}$.

Proof: See Sec. B.3. □

Next, our task is to show that the recursion in equation 1 and equation 2 satisfy a bound of the form $\Phi\left(\underline{\boldsymbol{w}}^{r}\right)+\theta\mathcal{D}_{r,0} \leq \upsilon^r \times \left(\Phi\left(\underline{\boldsymbol{w}}^{0}\right)+\theta\Phi\left(\underline{\boldsymbol{w}}^{0}\right)\right)$, for some $\theta > 0$ and $\upsilon \in (0,1)$, which is the desired result. In summary, we prove the bounds on the loss (in terms of drift) in Lemma 1 and the drift (in terms of the loss) in Lemma 2. Note that proving the above bound involves carefully handling the drift due to multiple local rounds. Using the above results, we constructed a Lypunov function in terms of drift and loss for some $\theta > 0$, and prove the following bound by appropriately choosing the learning rate η and constants:

Lemma 3. *By choosing* η *as in Theorem 1 for some* $\theta > 0$*, we obtain the following*

$$\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)+\theta\mathcal{D}_{r+1,0} \quad \leq \quad \left(1-\frac{\eta\mu}{16}\right)\left(\Phi\left(\underline{\boldsymbol{w}}^{r}\right)+\theta\mathcal{D}_{r,0}\right). \tag{3}$$

Proof: See Sec. B.4. □

The above Lemma shows that the drift and the loss go down to zero exponentially fast (linear convergence), and results in Theorem 1. The above approach is quite different from the existing work (Koloskova et al., 2020).

4 Server Setting: FedAvg with Improved Rates

In this section, we show that our analysis specialized to the *server* setting, i.e., $\lambda_2 = 0$ enables us to show that there exists an optimal number of local rounds T that maximizes the rate of convergence. In particular, we provide a closed form solution to the optimum number of local rounds, and show that an improved rate of $\mathcal{O}(\log(1/\epsilon))$ can be achieved as opposed to the existing work (Koloskova et al., 2020) where they show a convergence rate of $\mathcal{O}(T\log(T/\epsilon))$. Further, the same observation is made empirically in both *server* and

the *decentralized* settings (see Sec. 5).
Optimal local rounds: Consider the equation for drift from Lemma 2

$$\mathcal{D}_{r,0} \leq \left(\left(1+\frac{1}{\psi}\right)\lambda_2^2+\eta^2\lambda_2^4\beta T^2 L_{max}^2\right)\mathcal{D}_{r-1,0}+2\eta^2\lambda_2^4\beta T^2 L_{max} N\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right], \tag{4}$$

and a bound on the loss function from equation 35 of Sec. B.2

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \leq\ & \mathbb{E}\left[\left(\left(1-\frac{\eta\mu}{4}\right)^T+\frac{64\eta^4T^3l_{max}LL_{max}}{\mu_{min}}\right)\Phi\left(\underline{\boldsymbol{w}}^r\right)+\frac{4\eta TL^2}{N}\left\|(Q-P)W^{r,0}\right\|_F^2\right.\\ &+\left.\frac{2\eta TL^2}{N}\left[\left(\frac{16l_{max}\eta^2T^2L_{max}^2}{\mu_{min}}+4\lambda_2^2\eta^2\gamma L_{max}^2T^2\right)\mathcal{D}_{r,0}+4\eta^2\gamma T^2\lambda_2^2\left\|\partial\Phi\left(\underline{W}^{r,0}\right)\right\|_F^2\right]\right] \end{aligned} \tag{5}$$

It is evident from the above equations that the drift increases with T, as expected. However, a part of the expression in the average loss decreases with T (more specifically, the term $\left(1-\frac{\eta\mu}{4}\right)^T$) while the other terms increase with T. In principle, one should be able to characterize the optimal T. However, the above is a complicated expression to optimize with respect to T. To get more insights into the effect of T, in the following, we look at the *server* setting, which is a special case of our framework. The *server* setting consists of the central server which coordinates the information sharing among participating clients. We obtain the *server* setting by making the second largest eigenvalue of the mixing matrix, i.e., $\lambda_2 = 0$ in the *decentralized* case. Now using the fact that $\lambda_2 = 0$ in equation 4 and equation 5 lead to $\mathcal{D}_{r,0} = 0$ and

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \leq \left(\left(1-\frac{\eta\mu}{4}\right)^T+\frac{64\eta^3T^3l_{max}L^2L_{max}}{\mu_{min}}\right)\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^r\right)\right], \tag{6}$$

respectively. Choosing $\eta \leq \frac{64l_{max}L^2L_{max}}{\mu_{min}T}$, and utilizing the upper bound $e^{-x} \leq 1-x+\frac{x^2}{2}$, for all $x \geq 0$ in equation 6, result in

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \leq \left(1-\frac{\eta\mu T}{4}+\eta^2T^2\left(\frac{\mu^2}{8}+1\right)\right)\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^r\right)\right].$$

Now we choose T such that the right hand side above is minimized, i.e.,

$$\inf_T\left[1-\frac{\eta\mu T}{4}+\eta^2T^2\left(\frac{\mu^2}{8}+1\right)\right].$$

Note that the above is a convex function. Hence, differentiating the above w.r.t T and equating it to 0, we get

$$T = T_{th} = \frac{\mu}{\eta\left(\mu^2+8\right)}.$$

The above analysis leads to the following "faster" convergence rate for the *decentralized* setting:

Corollary 2. *By choosing the number of local updates such that $T = T_{th} = \frac{\mu}{\eta(\mu^2+8)}$, the iterates generated by Algorithm 1 achieve an ϵ accurate point after $R = \mathcal{O}\left(\frac{8(\mu^2+8)}{\mu^2}\log\left(\frac{\Phi(\underline{\boldsymbol{w}}^0)}{\epsilon}\right)\right)$ communication rounds.*

In the above analysis, we capture the effect of local updates on the performance of the decentralized FedAvg. Specifically, we show that there exists an optimal number of local updates T beyond which the convergence of the algorithm slows down and one needs to choose T carefully to achieve the optimal convergence guarantees. Again, this is the first result establishing linear convergence of FedAvg in the *decentralized* setting when minimizing non-convex functions satisfying PL-inequality in the interpolation regime. In the next section, we present the experimental results.

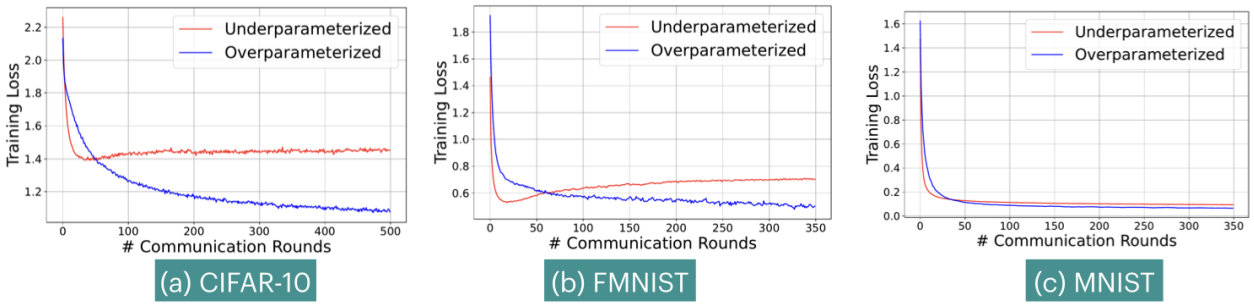

Figure 2: Training loss on different datasets versus the communication rounds for FedAvg in the *decentralized* setting.

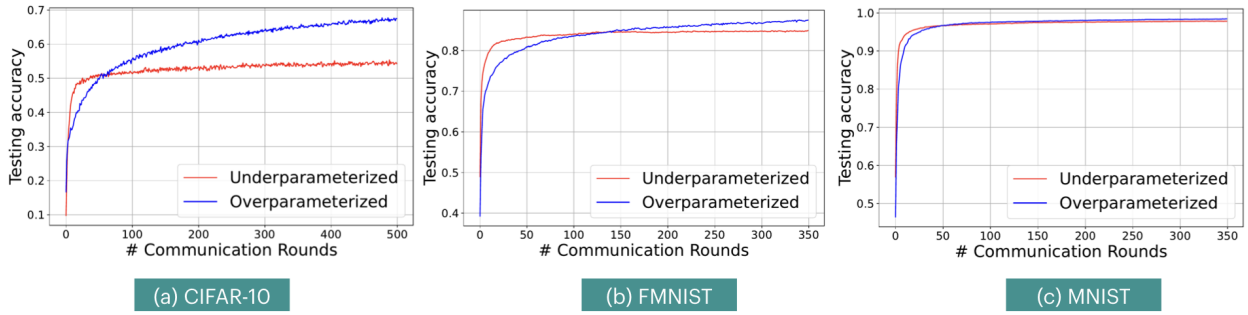

Figure 3: Testing accuracy on different datasets versus the communication rounds for FedAvg in the *decentralized* setting.

5 Experimental Evaluation

In this section, we experimentally validate our theoretical findings for the *decentralized* versions of FedAvg. First, we present the experimental setup for various settings.

5.1 Setup: *Decentralized* FedAvg

We use 60 edge devices to run the *Decentralized* FedAvg algorithm with multiple local SGD steps and then broadcast the updated model with the nodes connected to it. We consider the image classification tasks on CIFAR-10, MNIST, and FMNIST datasets using an overparameterized simple regression and Deep Neural Network (DNN) models. We have implemented all our experiments on NVIDIA *DGX A*100. The experimental setup consists of the following model and data set:

Overparameterized regression. Here we consider a simple regression model with 3 linear layers. There are 231490 trainable parameters with no activation function. We evaluate the performance of *Decentralized* FedAvg algorithm on the MNIST dataset for an image classification task.

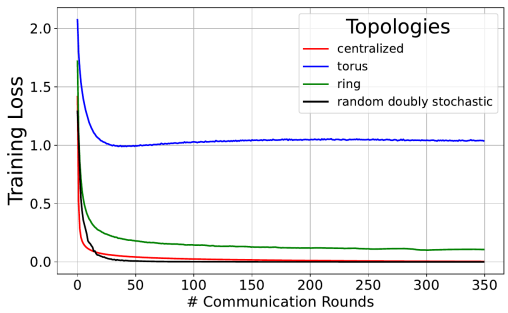

Figure 5: Training loss versus the communication rounds for FedAvg in the *decentralized* setting. Here, random doubly stochastic case has 5 clients while for others we have used 60 clients.

Deep neural network. We consider an image classification task under two different settings: underparameterized and overparameterized settings. In this case,

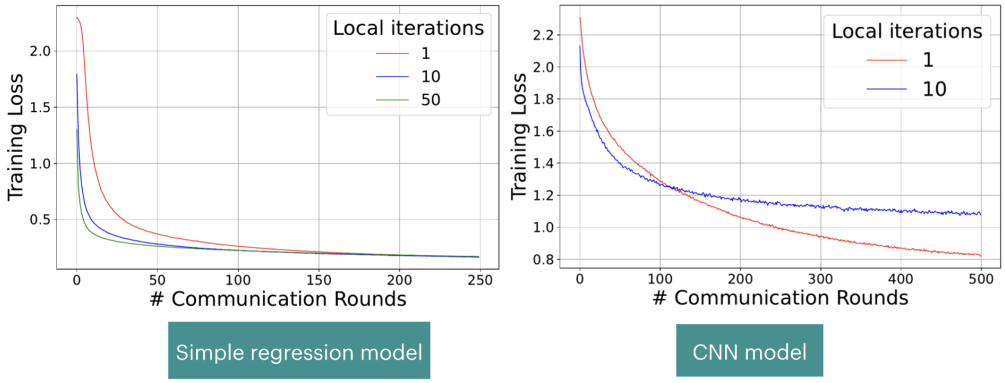

Figure 4: Effect of T on the convergence of *Decentralized* FedAvg for simple regression and CNN model.

each device implements a convolutional neural network (CNN) model. We consider the CIFAR-10, MNIST and FMNIST datasets. In the overparameterized setting, each edge device implements a three hidden layer convolutional neural network (CNN) with 256, 128, and 64 filters followed by three linear layers having 1642849 trainable parameters for CIFAR-10 and two linear layers for MNIST and FMNIST with 1046426 trainable parameters. On the contrary, the underparameterized setting considers a relatively smaller neural network. In this setting, each device implements two hidden layer CNN having 25 and 52 filters followed by two linear layers for CIFAR-10 and one linear layer for MNIST and FMNIST datasets. We set the number of local updates $T = 10$ and pick the tunable learning rate in the range $\eta \in [0.001 : 0.01]$ for CIFAR-10, MNIST, and FMNIST datasets. We consider that each device has 490 training samples and 90 test samples for the CIFAR-10 dataset. On the other hand, for MNIST and FMNIST datasets, 540 samples are used for training and 80 samples are used for testing.

In this setting, we run Algorithm 1 for the following networks (i) ring, (ii) random doubly stochastic, and (iii) torus topologies. For the *Decentralized* FedAvg, we compare (a) the performance of *Decentralized* FedAvg with both underparametrized and overparameterized neural network models, (b) effect of topology on the convergence, and (c) effect of local updates on the convergence. In the following, we provide a detailed experimental results.

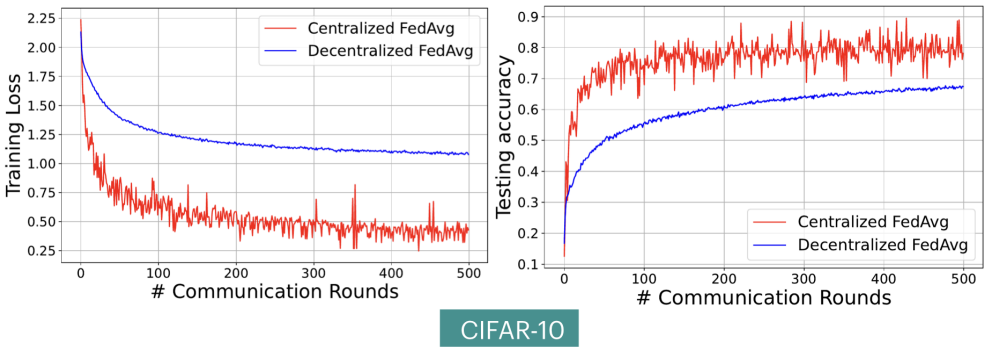

Figure 6: Training loss and Testing accuracy for centralized ($\lambda_2 = 0$) and *Decentralized* FedAvg algorithm with ring topology ($\lambda_2 = 0.33$) on CIFAR-10 dataset versus communication rounds.

5.2 Experimental Results for Different Settings

Using the settings described above, here we present experimental results for *Decentralized* FedAvg's, and corroborate various theoretical findings made in this paper:

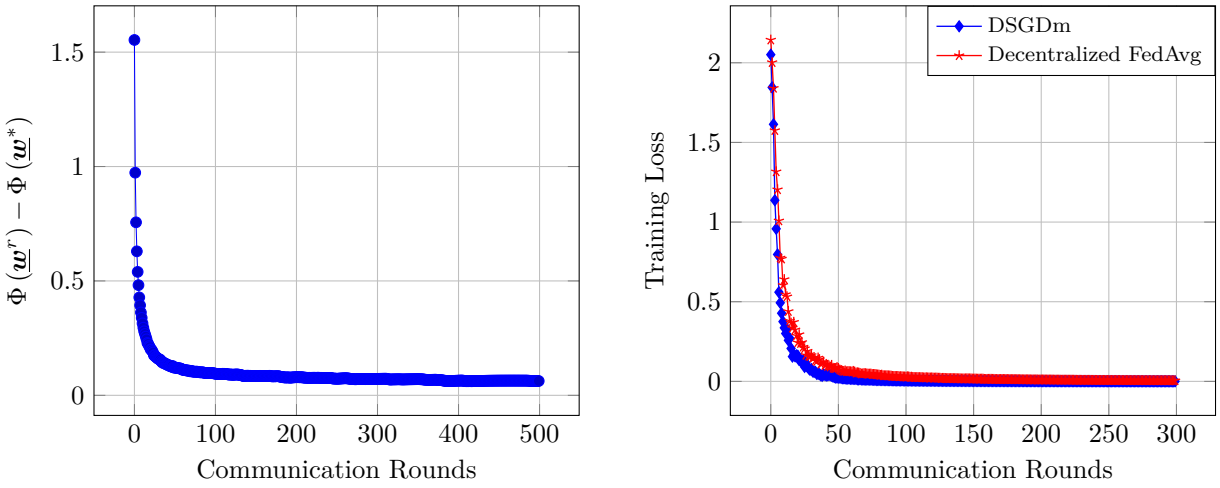

Figure 7: Optimality gap for *Decentralized* FedAvg algorithm with ring topology on MNIST dataset versus communication rounds. (see (a)) and Comparison of the DSGDm and *Decentralized* FedAvg. (see (b)).

1. **Underparameterized versus overparametrized.** Fig. 2 shows the plots of training loss of FedAvg in the *decentralized* setting for underparameterized and overparameterized models on MNIST, FMNIST, CIFAR-10 datasets. As established in Theorem 1, the loss of FedAvg in the *decentralized* setting diminishes rapidly for the overparameterized models compared to the underparameterized models. This is due to the fact that the PL inequality is satisfied for overparameterized systems which helps to reach the global optimum at a linear rate as demonstrated by Theorem 1. Fig. 3 show plots for testing accuracy for FedAvg in the *decentralized* setting. As expected the convergence speed of underparameterized case is slower than the overparameterized case.

2. **Effect of local updates T.** Fig. 4 shows plots of the training losses on MNIST dataset for the FedAvg under the *decentralized* setting on the overparameterized regression model and the CNN. From equation 6, we see that as T increases, the convergence speed either decreases or increases depending on the coefficient of T^3 in the second term. We capture this phenomenon in Fig. 4. In particular, as T increases, the rate of convergence increases for simple regression model while it decreases/saturates for the CNN based DNN model. One plausible explanation is that the smoothness constants of simple regression is small, and hence results in smaller second term in equation 6. However, for CNN based DNN, the second term dominates, and hence results in slower convergence with T.

3. **Effect of optimality gap.** Fig. 7 (a) shows the plot of the optimality gap, i.e., $\Phi(\underline{\boldsymbol{w}}^r) - \Phi(\underline{\boldsymbol{w}}^*)$ versus R. Here $\boldsymbol{w}^*$ is an approximate optimal solution obtained by running a centralized algorithm for a sufficient number of rounds. As expected the optimality gap decreases exponentially with R.

4. **Comparison with the existing work.** Fig. 7 (b) shows the comparison of Decentralized Stochastic Gradient Descent with momentum (DSGDm) (Lin et al., 2021) with the *Decentralized* FedAvg algorithm. We show the training loss versus R for overparameterized CNN model using MNIST dataset in both the cases. We see DSGDm outperforms the *Decentralized* FedAvg algorithm as expected.

5. **Comparison with different topologies in the *decentralized* case.** Fig. 5 shows the training loss versus the communication rounds R for overparameterized CNN model using MNIST dataset with $T = 10$ for four different topologies. Since server topology has $\lambda_2 = 0$, it outperforms the network with ring topology and a random (doubly) stochastic matrix. However, the torus topology does not satisfy the conditions required, i.e., symmetric and doubly stochastic matrix, and hence cannot be used for corroborating our theoretical findings. Nevertheless, we have conducted experiments with torus topology, and Fig. 5 shows that the torus has the worst convergence performance. One reason for this could be that the ring topology has more structure, i.e., it has a symmetric and doubly stochastic mixing matrix P as

opposed to the torus topology. The theoretical analysis of networks with general topology is relegated to our future work.

6 Conclusion

In this work, we performed a theoretical analysis of the well known FedAvg algorithm for the class of smooth non-convex overparameterized systems in the interpolation regime. We considered the *decentralized* setting where nodes communicate over an undirected graph. In this regime, it is well know that neural networks with non-convex loss functions typically satisfy an inequality called Polyak-Lojasiewicz (PL) condition. Assuming PL condition, we showed that the FedAvg algorithm achieves linear convergence rate $\mathcal{O}(T^2 \log(1/\epsilon))$, where ϵ is the desired solution accuracy, and T is the number of local SGD updates at each node. As opposed to the standard analysis of the FedAvg algorithm, we showed that our approach does not require bounded heterogeneity, variance, and gradient assumptions. We captured the heterogeneity in FL training through sample-wise and local smoothness of loss functions. Finally, we carried out experiments on multiple real-world datasets to confirm our theoretical observations.

Acknowledgements: This work is in part supported by SERB-CRG (grant number: CRG/2021/007502) and SERB-MATRICS (grant number: MTR/2021/000575). We thank the anonymous reviewers for their helpful comments which have improved the paper.

References

Sai Aparna Aketi, Abolfazl Hashemi, and Kaushik Roy. Global update tracking: A decentralized learning algorithm for heterogeneous data. *Advances in Neural Information Processing Systems*, 36, 2024.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

Aleksandr Beznosikov, Pavel Dvurechenskii, Anastasiia Koloskova, Valentin Samokhin, Sebastian U Stich, and Alexander Gasnikov. Decentralized local stochastic extra-gradient for variational inequalities. *Advances in Neural Information Processing Systems*, 35:38116–38133, 2022.

Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Local sgd with periodic averaging: Tighter analysis and adaptive synchronization. *Adv. Neural Inf. Process.*, 32, 2019.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Yifan Hua, Kevin Miller, Andrea L Bertozzi, Chen Qian, and Bao Wang. Efficient and reliable overlay networks for decentralized federated learning. *SIAM Journal on Applied Mathematics*, 82(4):1558–1586, 2022.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *ECML PKDD*, pp. 795–811. Springer, 2016.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *Int. Conf. on ML*, pp. 5132–5143. PMLR, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. First analysis of local gd on heterogeneous data. *arXiv preprint arXiv:1909.04715*, 2019.

Anastasia Koloskova, Tao Lin, Sebastian U Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. *arXiv preprint arXiv:1907.09356*, 2019.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *Int. Conf. on ML*, pp. 5381–5393. PMLR, 2020.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019a.

Xiang Li, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Communication-efficient local decentralized sgd methods. *arXiv preprint arXiv:1910.09126*, 2019b.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Tao Lin, Sai Praneeth Karimireddy, Sebastian U Stich, and Martin Jaggi. Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data. *arXiv preprint arXiv:2102.04761*, 2021.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. *arXiv preprint arXiv:2003.00307*, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022a.

Wei Liu, Li Chen, and Wenyi Zhang. Decentralized federated learning: Balancing communication and computing costs. *IEEE Transactions on Signal and Information Processing over Networks*, 8:131–143, 2022b.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *Int. Conf. on ML*, pp. 3325–3334. PMLR, 2018.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and stats.*, pp. 1273–1282. PMLR, 2017.

Giorgi Nadiradze, Amirmojtaba Sabour, Peter Davies, Shigang Li, and Dan Alistarh. Asynchronous decentralized sgd with quantized and local updates. *Advances in Neural Information Processing Systems*, 34: 6829–6842, 2021.

Quynh Nguyen, Marco Mondelli, and Guido F Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *Int. Conf. on ML*, pp. 8119–8129. PMLR, 2021.

Quynh N Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. *Adv. Neural Inf. Process.*, 33:11961–11972, 2020.

Zhaonan Qu, Kaixiang Lin, Jayant Kalagnanam, Zhaojian Li, Jiayu Zhou, and Zhengyuan Zhou. Federated learning's blessing: Fedavg has linear speedup. *arXiv preprint arXiv:2007.05690*, 2020.

Tomer Sery, Nir Shlezinger, Kobi Cohen, and Yonina C. Eldar. Over-the-air federated learning from heterogeneous data. *IEEE Transactions on Signal Processing*, 69:3796–3811, 2021. doi: 10.1109/TSP.2021.3090323.

Bingqing Song, Ioannis Tsaknakis, Chung-Yiu Yau, Hoi To Wai, and Mingyi Hong. Distributed optimization for overparameterized problems: Achieving optimal dimension independent communication complexity. In *Advances in Neural Information Processing Systems*.

Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Tao Sun, Dongsheng Li, and Bao Wang. Decentralized federated averaging. *arXiv preprint arXiv:2104.11375*, 2021.

Yuki Takezawa, Han Bao, Kenta Niwa, Ryoma Sato, and Makoto Yamada. Momentum tracking: Momentum acceleration for decentralized deep learning on heterogeneous data. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=8koy8QuTZD`.

Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of local-update sgd algorithms. *Journal of Machine Learning Research*, 22(213):1–50, 2021. URL `http://jmlr.org/papers/v22/20-147.html`.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. A novel framework for the analysis and design of heterogeneous federated learning. *IEEE Transactions on Signal Processing*, 69:5234–5249, 2021. doi: 10.1109/TSP.2021.3106104.

Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6):1205–1221, 2019.

Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? In *Int. Conf. on ML*, pp. 10334–10343. PMLR, 2020a.

Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020b.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*, 2021.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *Int. Conf. on ML*, pp. 7184–7193. PMLR, 2019a.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5693–5700, 2019b.

Anastasia Koloskova Yue Liu, Tao Lin and Sebastian U. Stich. Decentralized gradient tracking with local steps. *Optimization Methods and Software*, 0(0):1–28, 2024. doi: 10.1080/10556788.2024.2322095. URL `https://doi.org/10.1080/10556788.2024.2322095`.

Miaoxi Zhu, Li Shen, Bo Du, and Dacheng Tao. Stability and generalization of the decentralized stochastic gradient descent ascent algorithm. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL `https://openreview.net/forum?id=ql6LVyi2Dg`.

Tongtian Zhu, Fengxiang He, Kaixuan Chen, Mingli Song, and Dacheng Tao. Decentralized sgd and average-direction sam are asymptotically equivalent. In *International Conference on Machine Learning*, pp. 43005–43036. PMLR, 2023b.

A Appendix

A.1 Compact Notations and Useful Lemmas

We simplify the presentation of the proof by using the following matrix notations. Let the local average of the parameters be denoted by $\underline{W}_l^r := [\underline{\boldsymbol{w}}_1^r, \underline{\boldsymbol{w}}_2^r, \ldots, \underline{\boldsymbol{w}}_N^r]^T \in \mathbb{R}^{N\times d}$, where $\underline{\boldsymbol{w}}_k^r \in \mathbb{R}^d$ is the parameter vector at node k. The `Aggregation` step of Algorithm 1 can be compactly written in matrix form as

$$\underline{\boldsymbol{w}}_k^{r+1} = \sum_{i\in\mathcal{N}_k} p_{k,i}\boldsymbol{w}_k^{r,T} \quad \equiv \quad \underline{W}_l^{r+1} = PW^{r,T}, \tag{7}$$

where $\mathcal{N}_k := \{i : p_{k,i} > 0\}$, and the symbol $\equiv$ means "equivalent to". Further, we define the global average as

$$\underline{\boldsymbol{w}}^r := \frac{1}{N}\sum_{k=1}^{N} \underline{\boldsymbol{w}}_k^r \quad \equiv \quad \underline{W}^r = Q\underline{W}_l^r, \tag{8}$$

where $Q := \frac{1}{N}\mathbf{1}\mathbf{1}^T$. Now, let us represent the gradients compactly in the matrix form as

$$\partial\hat{\Phi}\left(W^{r,t}\right) := \left[\frac{1}{b}\sum_{j\in B_1^{r,t}} G_{1,j}^{(r,t)}, \frac{1}{b}\sum_{j\in B_2^{r,t}} G_{2,j}^{(r,t)}, \ldots, \frac{1}{b}\sum_{j\in B_N^{r,t}} G_{N,j}^{(r,t)}\right], \tag{9}$$

where $G_{l,j}^{(r,t)} := \nabla\Phi_{l,j}\left(\boldsymbol{w}_l^{r,t}\right)$. The mixing matrix P also preserves the average, and hence $QP = P$. In the following subsection, we provide a Lemma that relates the local average with the drift. Next, we present two Lemmas that will be used in proving the convergence result of *Decentralized* FedAvg algorithm, in particular, while bounding the global drift.

Lemma 4. *(See (Horn & Johnson, 2012)) For any matrices $A \in \mathbb{C}^{N\times N}$ and $B \in \mathbb{C}^{N\times d}$, we have $\|AB\|_F^2 \leq \|A\|_{op}^2 \|B\|_F^2$, where $\|A\|_{op}$ denotes the operator norm of A.*

Lemma 5. *(See Lemma 1 in (Sun et al., 2021)) Suppose Assumption 5 holds, then for any $m \in \mathbb{N}$, the mixing matrix P satisfies $\|P^m - Q\|_{op} \leq \lambda_2^m$, where λ_2 is the second largest eigenvalue of the mixing matrix P, and $Q := \frac{1}{N}\mathbf{1}\mathbf{1}^T$.*

The above Lemmas are standard which come in handy while bounding the consensus error (Koloskova et al., 2020; Liu et al., 2022b; Wang & Joshi, 2021). See Liu et al. (2022b); Wang & Joshi (2021); Sun et al. (2021) for the detailed proofs.

B Proof of Theorem 1

In this section, we will prove the main theorem by proving Lemmas 1 and 2. The proof mainly consists of two intermediate steps, namely bounding i) the local loss (see Lemma 1) using L_k smoothness (see Definition 1) and local PL inequality to show that the loss at local parameter is bounded in terms of the loss at the global average parameter and the drift and ii) the global drift (see Lemma 2).

B.1 Useful Lemma to Prove Theorem 1

Lemma 1. *The expected local loss function $\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)$ satisfies the following bound*

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \quad \leq \quad \frac{2L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r - \underline{\boldsymbol{w}}^r\right\|_2^2 + \frac{2}{\mu_{min}}\mathbb{E}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|^2, \tag{10}$$

where $\mu_{min} := \min_{k\in[N]}\{\mu_k\}$.

Proof: Using Assumption 1, we have

$$\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right) \leq \Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right) + \left\langle \nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right), \boldsymbol{w}_k^{r,\tau} - \boldsymbol{w}_k^{r,\tau-1}\right\rangle + \frac{L_k}{2}\left\|\boldsymbol{w}_k^{r,\tau} - \boldsymbol{w}_k^{r,\tau-1}\right\|_2^2. \tag{11}$$

We know from Step 7 of the **Algorithm** 1 that $\boldsymbol{w}_k^{r,\tau} - \boldsymbol{w}_k^{r,\tau-1} = -\frac{\eta}{b}\sum_{j\in B_k^{r,\tau-1}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)$. Using this in equation 11, we get

$$\begin{aligned}
\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right) &\leq \Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right) - \eta\left\langle \nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right), \frac{1}{b}\sum_{j\in B_k^{r,\tau-1}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\rangle + \frac{\eta^2 L_k}{2} G_k(r,\tau). \\
&= \Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right) - \eta\left\langle \nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right), \frac{1}{b}\sum_{j\in B_k^{r,\tau-1}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1)}\right)\right\rangle \\
&+ \frac{\eta^2 L_k}{2b^2}\sum_{j\in B_k^{r,\tau-1}}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2 + \frac{\eta^2 L_k}{2b^2}\sum_{j\neq j'}\left\langle \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right), \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\rangle. \\
&\leq \Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right) - \eta\left\langle \nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right), \frac{1}{b}\sum_{j\in B_k^{r,\tau-1}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1)}\right)\right\rangle \\
&+ \frac{\eta^2 L_{max}}{2b^2}\sum_{j\in B_k^{r,\tau-1}}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2 + \frac{\eta^2 L_{max}}{2b^2}\sum_{j\neq j'}\left\langle \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right), \nabla\Phi_{k,j'}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\rangle
\end{aligned} \tag{12}$$

where $G_k(r,\tau) := \left\|\frac{1}{b}\sum_{j\in B_k^{r,\tau-1}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2$, and $L_{max} := \max_k L_k$. Taking expectation of the above leads to

$$\begin{aligned}
\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] &\leq \mathbb{E}\Bigg[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right) - \eta\left\langle \nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right), \nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\rangle + \frac{\eta^2 L_{max}}{2b}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2 \\
&+ \frac{\eta^2 L_{max} b(b-1)}{2b^2}\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2\Bigg].
\end{aligned}$$

Applying smoothness assumption of each sample, i.e., $\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2 \leq 2l_{k,j}\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)$, we have

$$\begin{aligned}
\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] &\leq \mathbb{E}\Bigg[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right) - \eta\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2 + \frac{\eta^2 L_{max} l_{k,j}}{b}\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right) \\
&+ \frac{\eta^2 L_{max} b(b-1) L_k}{b^2}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right]\Bigg]. \\
&\leq \Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right) - \eta\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2 + \frac{\eta^2 L_{max} l_{max}}{b}\mathbb{E}\left[\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right] \\
&+ \frac{\eta^2 L_{max}^2 b(b-1)}{b^2}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right],
\end{aligned} \tag{13}$$

where $l_{max} := \max_k L_k$. From the local PL inequality (see definition 2), it follows that $\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2 \geq \mu_{min}\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)$ for $k = \{1, 2, \ldots, N\}$, where $\mu_{min} := \min_{k\in[N]}\{\mu_k\}$. Using this in equation 13 results in

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left[1 - \eta\mu_{min} + \eta^2\left(\frac{l_{max}L_{max}}{b} + \frac{L_{max}^2 b(b-1)}{b^2}\right)\right]\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right].$$

By setting $\eta \leq \frac{\mu_{min}}{2\left[\frac{l_{max}L_{max}}{b} + \frac{L_{max}^2 b(b-1)}{b^2}\right]}$, the above can be further bounded as

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left(1 - \frac{\eta\mu_{min}}{2}\right)\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right].$$

Since $\boldsymbol{w}_k^{r,0} = \underline{\boldsymbol{w}}_k^r$, the above can be written as

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left(1 - \frac{\eta\mu_{min}}{2}\right)^{\tau} \mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}_k^r\right)\right]. \tag{14}$$

Using the local PL inequality, i.e., $\Phi_k\left(\underline{\boldsymbol{w}}_k^r\right) \leq \frac{1}{\mu_{min}} \left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}_k^r\right)\right\|_2^2$ in equation 14, we have

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left(1 - \frac{\eta\mu_{min}}{2}\right)^{\tau} \frac{1}{\mu_{min}} \mathbb{E}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}_k^r\right)\right\|_2. \tag{15}$$

Now, adding and subtracting the term $\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)$ in the above, and using the fact that $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we get

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left(1 - \frac{\eta\mu_{min}}{2}\right)^{\tau} \frac{2}{\mu_{min}} \mathbb{E}\left(\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}_k^r\right) - \nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|_2^2 + \left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|_2^2\right).$$

Using L_k smoothness assumption (see Assumption 3), we have

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left(1 - \frac{\eta\mu_{min}}{2}\right)^{\tau} \mathbb{E}\left(\frac{2L_k^2}{\mu_{min}} \left\|\underline{\boldsymbol{w}}_k^r - \underline{\boldsymbol{w}}^r\right\|_2^2 + \frac{2}{\mu_{min}} \left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|_2^2\right).$$

Choosing $\eta < \frac{2}{\mu_{min}}$ and using the fact that $L_{max} = \max_k L_k$, we get

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \quad \leq \quad \frac{2L_{max}^2}{\mu_{min}} \mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r - \underline{\boldsymbol{w}}^r\right\|_2^2 + \frac{2}{\mu_{min}} \mathbb{E}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|^2.$$

□

Using smoothness condition, the above leads to the following corollary. The below result comes in handy while proving the main result.

Corollary 1. *The function $\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)$ satisfies local PL inequality and can be bounded in terms of global average parameter i.e., $\Phi_k\left(\underline{\boldsymbol{w}}^r\right)$ as follows*

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \quad \leq \quad \frac{2L_{max}^2}{\mu_{min}} \mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r - \underline{\boldsymbol{w}}^r\right\|_2^2 + \frac{4L_{max}}{\mu_{min}} \mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right], \tag{16}$$

where $\mu_{min} := \min_{k\in[N]}\{\mu_k\}$ and $L_{max} := \max_k L_k$.

B.2 Proof of Lemma 1

From L-smoothness assumption (see 1) of $\Phi\left(\boldsymbol{w}\right)$, we have

$$\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right) \quad \leq \quad \Phi\left(\underline{\boldsymbol{w}}^{r,t}\right) + \langle\nabla\Phi(\underline{\boldsymbol{w}}^{r,t}), \underline{\boldsymbol{w}}^{r,t+1} - \underline{\boldsymbol{w}}^{r,t}\rangle + \frac{L}{2}\left\|\underline{\boldsymbol{w}}^{r,t+1} - \underline{\boldsymbol{w}}^{r,t}\right\|^2. \tag{17}$$

Using step 7 of **Algorithm 2**, we have $\boldsymbol{w}_i^{r,t+1} = \boldsymbol{w}_i^{r,t} - \frac{\eta}{b}\sum_{j\in B_i^{r,t}} \nabla\Phi_{i,j}\left(\boldsymbol{w}_i^{r,t}\right)$. Multiplying both sides by $p_{k,i}$ and summing over $i \in \mathcal{N}_k$, we get

$$\underline{\boldsymbol{w}}_k^{r,t+1} \quad = \quad \underline{\boldsymbol{w}}_k^{r,t} - \frac{\eta}{b}\sum_{i\in\mathcal{N}_k} p_{k,i} \sum_{j\in B_i^{r,t}} \nabla\Phi_{i,j}\left(\boldsymbol{w}_i^{r,t}\right). \tag{18}$$

Averaging on both sides over $k \in [N]$, we get

$$\underline{\boldsymbol{w}}^{r,t+1} \quad = \quad \underline{\boldsymbol{w}}^{r,t} - \frac{\eta}{bN}\sum_{k=1}^{N} \sum_{j\in B_k^{r,t}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right).$$

Substituting for $\underline{\boldsymbol{w}}^{r,t+1} - \underline{\boldsymbol{w}}^{r,t}$ from the above udpate in equation 17, we get

$$\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right) \quad \leq \quad \Phi\left(\underline{\boldsymbol{w}}^{r,t}\right) - \eta \left\langle \nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right), \frac{1}{bN}\sum_{k=1}^{N}\sum_{j\in B_k^{r,t}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\right\rangle + \frac{\eta^2 L}{2b^2N^2}\left\|\mathcal{G}^{r,t}\right\|^2,$$

where $\mathcal{G}^{r,t} := \sum_{k=1}^{N}\sum_{j\in B_k^{r,t}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)$. Taking expectation conditioning on $\boldsymbol{w}_k^{r,t}$ and past, we get

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] \quad \leq \quad & \mathbb{E}\Bigg[\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right) - \eta \underbrace{\left\langle \nabla\Phi(\underline{\boldsymbol{w}}^{r,t}), \frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\right\rangle}_{:=\mathcal{A}_1} + \frac{\eta^2 L}{2}\Bigg(\underbrace{\frac{1}{b^2N^2}\sum_{k=1}^{N}\left\|\sum_{j\in B_k^{r,t}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2}_{:=\mathcal{A}_2} \\ & + \underbrace{\frac{1}{b^2N^2}\sum_{k\neq k'}\left\langle \sum_{j\in B_k^{r,t}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right), \sum_{i\in B_k^{r,t}}\nabla\Phi_{k',i}\left(\boldsymbol{w}_{k'}^{r,t}\right)\right\rangle}_{:=\mathcal{A}_3}\Bigg)\Bigg], \end{aligned} \tag{19}$$

First, consider the second term above, i.e., $\mathcal{A}_2$

$$\mathcal{A}_2 \quad = \quad \frac{1}{b^2N^2}\sum_{k=1}^{N}\sum_{j\in B_k^{r,t}}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2 + \frac{1}{b^2N^2}\sum_{k=1}^{N}\sum_{j\neq j'}\left\langle \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right), \nabla\Phi_{k,j'}\left(\boldsymbol{w}_k^{r,t}\right)\right\rangle.$$

Taking expectation, we get

$$\mathbb{E}[\mathcal{A}_2] \quad = \quad \frac{1}{bN^2}\sum_{k=1}^{N}\mathbb{E}\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\|^2 + \frac{b(b-1)}{b^2N^2}\sum_{k=1}^{N}\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2. \tag{20}$$

Similarly the term $\mathcal{A}_3$ in equation 19 can be bounded by taking expectation as follows

$$\begin{aligned}\mathbb{E}[\mathcal{A}_3] \quad &= \quad \frac{1}{N^2}\sum_{k\neq k'}\left\langle \nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right), \nabla\Phi_{k'}\left(\boldsymbol{w}_{k'}^{r,t}\right)\right\rangle \\ &\overset{(a)}{\leq} \quad \frac{1}{2N^2}\sum_{k\neq k'}\left[\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2 + \|\nabla\Phi_{k'}\left(\boldsymbol{w}_{k'}^{r,t}\right)\|^2\right] \\ &= \quad \frac{2(N-1)}{2N^2}\sum_{k=1}^{N}\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2 \\ &\leq \quad \frac{1}{N}\sum_{k=1}^{N}\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2, \end{aligned} \tag{21}$$

where (a) follows from $\langle a, b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$. Next, we lower bound the term $\mathcal{A}_1$ in equation 19 as

$$\begin{aligned}\mathcal{A}_1 \quad &= \quad \frac{1}{2}\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\|^2 + \frac{1}{2}\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2 - \frac{1}{2}\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right) - \nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\|^2 \\ &\overset{\texttt{Jensen + smoothness}}{\geq} \quad \frac{1}{2}\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\|^2 + \frac{1}{2}\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2 - \frac{L^2}{2N}\sum_{k=1}^{N}\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\|^2. \end{aligned} \tag{22}$$

Substituting equation 20, equation 21 and equation 22 in equation 17, we get the following

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] \quad \leq \quad & \mathbb{E}\Bigg[\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right) - \frac{\eta}{2}\left\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 - \frac{\eta}{2}\left\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2 + \frac{\eta L^2}{2N}\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 \\ & + \frac{\eta^2 L}{2bN^2}\underbrace{\sum_{k=1}^{N}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2}_{:=\mathcal{A}_4} + \left(\frac{\eta^2 Lb(b-1)}{2b^2N^2} + \frac{\eta^2 L}{2N}\right)\underbrace{\sum_{k=1}^{N}\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2}_{:=\mathcal{A}_5}\Bigg]. \end{aligned} \tag{23}$$

The term $\mathcal{A}_4$ in equation 23 is bounded as follows

$$\begin{aligned}
\mathcal{A}_4 &\overset{(a)}{\leq} \sum_{k=1}^{N} 2\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)-\nabla\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2+\sum_{k=1}^{N} 2\left\|\nabla\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 \\
&\overset{(b)}{\leq} 2\sum_{k=1}^{N} l_{k,j}^2\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2+4\sum_{k=1}^{N} l_{k,j}\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right) \\
&\overset{(c)}{\leq} 2l_{max}^2\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2+4l_{max}\sum_{k=1}^{N}\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right),
\end{aligned}$$

where (a) follows by adding and subtracting the term $\nabla\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right)$ and using the fact that, $\|a+b\|^2 \leq 2\|a\|^2+2\|b\|^2$, (b) follows from Assumption 3, and (c) follows from the fact that $l_{max} := \max_{k,j} l_{k,j}$. Taking expectation, we get

$$\mathbb{E}\left[\mathcal{A}_4\right] \leq 2l_{max}^2\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2+4l_{max}\sum_{k=1}^{N}\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right)\right]. \tag{24}$$

The term $\mathcal{A}_5$ in equation 23 is bounded as

$$\begin{aligned}
\mathcal{A}_5 &\overset{(a)}{\leq} 2\sum_{k=1}^{N}\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)-\nabla\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2+2\sum_{k=1}^{N}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 \\
&\overset{(b)}{\leq} 2\sum_{k=1}^{N} L_k^2\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2+4\sum_{k=1}^{N} L_k\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right) \\
&\overset{(c)}{\leq} 2L_{max}^2\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2+4L_{max}\sum_{k=1}^{N}\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right),
\end{aligned} \tag{25}$$

where (a) follows by adding and subtracting $\nabla\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right)$, and (b) follows from Assumption 3 and (c) follows from $L_{max} := \max_k L_k$. Substituting upper bounds from equation 24 and equation 25 in equation 23, we get

$$\begin{aligned}
\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] \leq\ & \mathbb{E}\Bigg[\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)-\frac{\eta}{2}\left\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2-\frac{\eta}{2}\left\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2 \\
&+\left(\frac{\eta L^2}{2N}+\frac{\eta^2 L l_{max}^2}{bN^2}+\frac{\eta^2 L L_{max}^2}{N^2}+\frac{\eta^2 L L_{max}^2}{N}\right)\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2 \\
&+\left(\frac{2\eta^2 L l_{max}}{bN}+\frac{2\eta^2 L L_{max}}{N}+2\eta^2 L L_{max}\right)\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\Bigg].
\end{aligned} \tag{26}$$

Now, using PL inequality (see definition 2), i.e., $\|\nabla\Phi(\boldsymbol{w})\|^2 \geq \mu\Phi(\boldsymbol{w})$, $\forall \boldsymbol{w}\in\mathbb{R}^d$ and rearranging, we get

$$\begin{aligned}
\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] \leq\ & \mathbb{E}\Bigg[\left(1-\frac{\eta\mu}{2}+\left(\frac{2\eta^2 L l_{max}}{bN}+\frac{2\eta^2 L L_{max}}{N}+2\eta^2 L L_{max}\right)\right)\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right) \\
&+\left(\frac{\eta L^2}{2N}+\frac{\eta^2 L l_{max}^2}{bN^2}+\frac{\eta^2 L L_{max}^2}{N^2}+\frac{\eta^2 L L_{max}^2}{N}\right)\frac{1}{N}\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2\Bigg].
\end{aligned}$$

Choosing $\eta \leq \min\left\{\frac{\mu}{4\left(\frac{2Ll_{max}}{bN}+\frac{2LL_{max}}{N}+2LL_{max}\right)}, \frac{L^2}{2\left(\frac{Ll_{max}^2}{bN}+\frac{LL_{max}^2}{N}+LL_{max}^2\right)}\right\}$, the above can be further bounded as

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] \leq \left(1-\frac{\eta\mu}{4}\right)\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\right]+\frac{\eta L^2}{N}\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2 \tag{27}$$

$$\overset{(a)}{\leq} \left(1-\frac{\eta\mu}{4}\right)\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\right]+\frac{2\eta L^2}{N}\sum_{k=1}^{N}\mathbb{E}\left(\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|^2+\left\|\underline{\boldsymbol{w}}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2\right), \tag{28}$$

In the above, (a) follows by adding and subtracting the term $\underline{\boldsymbol{w}}_k^{r,t}$ and using the fact that, $\|a+b\|^2 \leq 2\|a\|^2+2\|b\|^2$. First, let us consider the local drift term i.e., $\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|$ in equation 28. Telescoping the update from step 7 of **Algorithm** 1 we have,

$$\boldsymbol{w}_k^{r,t} = \boldsymbol{w}_k^{r,0}-\frac{\eta}{b}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right). \tag{29}$$

Further, consider the local average at node k, i.e., $\underline{\boldsymbol{w}}_k^{r,t}$

$$\underline{\boldsymbol{w}}_k^{r,t}=\sum_{i\in\mathcal{N}_k}p_{k,i}\boldsymbol{w}_i^{r,t}=\underline{\boldsymbol{w}}_k^{r,0}-\frac{\eta}{b}\sum_{\tau=0}^{t-1}\sum_{i\in\mathcal{N}_k}p_{k,i}\sum_{j\in\mathcal{B}_i^{r,\tau}}\nabla\Phi_{i,j}\left(\boldsymbol{w}_i^{r,\tau}\right). \tag{30}$$

Now noting the fact that $\boldsymbol{w}_k^{r,0}=\underline{\boldsymbol{w}}_k^{r,0}$ and using equation 29 and equation 30, we can bound the drift term as

$$\begin{aligned}
\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|^2 &= \sum_{k=1}^{N}\mathbb{E}\left\|\frac{\eta}{b}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)-\frac{\eta}{b}\sum_{\tau=0}^{t-1}\sum_{i\in\mathcal{N}_k}p_{k,i}\sum_{j\in\mathcal{B}_i^{r,\tau}}\nabla\Phi_{i,j}\left(\boldsymbol{w}_i^{r,\tau}\right)\right\|^2\\
&\overset{(a)}{\leq} 2\sum_{k=1}^{N}\mathbb{E}\left[\left\|\frac{\eta}{b}\sum_{\tau=0}^{t-1}G_{kj}(r,\tau)\right\|^2+\left\|\frac{\eta}{b}\sum_{\tau=0}^{t-1}\sum_{i\in\mathcal{N}_k}p_{k,i}G_{ij}(r,\tau)\right\|^2\right]\\
&\overset{(b)}{\leq} 2\sum_{k=1}^{N}\mathbb{E}\left[\frac{\eta^2 t}{b^2}\sum_{\tau=0}^{t-1}\|G_{kj}(r,\tau)\|^2+\frac{\eta^2 t}{b^2}\sum_{\tau=0}^{t-1}\left\|\sum_{i\in\mathcal{N}_k}p_{k,i}G_{ij}(r,\tau)\right\|^2\right],
\end{aligned}$$

where $G_{ij}(r,\tau):=\sum_{j\in\mathcal{B}_i^{r,\tau}}\nabla\Phi_{i,j}\left(\boldsymbol{w}_i^{r,\tau}\right)$. In the above, (a) follows from the fact that, $\|a+b\|^2\leq 2\|a\|^2+2\|b\|^2$, and (b) follows from the fact that for any vector $\boldsymbol{z}_i$, $\left(\sum_{i=1}^{N}\boldsymbol{z}_i\right)^2\leq N\sum_{i=1}^{N}(\boldsymbol{z}_i)^2$. The second term in (b) can be further bounded using Jensen's inequality as follows

$$\begin{aligned}
\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|^2 &\leq 2\sum_{k=1}^{N}\mathbb{E}\left[\frac{\eta^2 t}{b^2}\sum_{\tau=0}^{t-1}\|G_{kj}(r,\tau)\|^2+\frac{\eta^2 t}{b^2}\sum_{\tau=0}^{t-1}\sum_{i\in\mathcal{N}_k}p_{k,i}\|G_{ij}(r,\tau)\|^2\right]\\
&\leq 2\sum_{k=1}^{N}\mathbb{E}\left[\frac{\eta^2 t}{b}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}\left\|g_{kj}^{r,\tau}\right\|^2+\frac{\eta^2 t}{b}\sum_{\tau=0}^{t-1}\sum_{i\in\mathcal{N}_k}p_{k,i}\sum_{j\in\mathcal{B}_i^{r,\tau}}\|g_{ij}^{r,\tau}\|^2\right]\\
&\overset{(a)}{\leq} 2\sum_{k=1}^{N}\mathbb{E}\left[\frac{\eta^2 t}{b}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}2l_{k,j}\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)+\frac{\eta^2 t}{b}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}\sum_{i\in\mathcal{N}_k}p_{k,i}2l_{i,j}\Phi_{i,j}\left(\boldsymbol{w}_i^{r,\tau}\right)\right]\\
&\overset{(b)}{=} \mathbb{E}\left[\frac{2\eta^2 t}{b}\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}2l_{max}\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)+\frac{2\eta^2 t}{b}\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}\sum_{i\in\mathcal{N}_k}p_{k,i}2l_{max}\Phi_{i,j}\left(\boldsymbol{w}_i^{r,\tau}\right)\right],
\end{aligned}$$

where $g_{kj}^{r,\tau} := \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)$. In the above, (a) follows from smoothness assumption and (b) follows from the fact that the mixing matrix P preserves the average and $l_{max} := \max_{k,j} l_{k,j}$. Simplifying the above results in

$$\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|^2 \quad\leq\quad \mathbb{E}\left[\frac{8\eta^2 t l_{max}}{b}\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)\right].$$

Taking expectation, we get

$$\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|^2 \quad\leq\quad 8\eta^2 t l_{max}\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right]. \tag{31}$$

From equation 16 of Corollary 1, we have, $\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \frac{2L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r-\underline{\boldsymbol{w}}^r\right\|_2^2+\frac{4L_{max}}{\mu_{min}}\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right]$. Using this in equation 31, we get

$$\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|^2 \quad\leq\quad 8\eta^2 t l_{max}\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\left(\frac{2L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r-\underline{\boldsymbol{w}}^r\right\|_2^2+\frac{4L_{max}}{\mu_{min}}\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right]\right).$$

Simplifying the above results in

$$\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|^2 \quad\leq\quad 16\eta^2 t^2 l_{max}L_{max}^2\sum_{k=1}^{N}\frac{\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r-\underline{\boldsymbol{w}}^r\right\|_2^2}{\mu_{min}}+32\eta^2 t^2 l_{max}L_{max}\sum_{k=1}^{N}\frac{\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right]}{\mu_{min}}. \tag{32}$$

Next, let us consider the global drift term i.e., $\sum_{k=1}^{N}\left\|\underline{\boldsymbol{w}}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|_2^2$ in equation 28, which can be rewritten in matrix notation as $\mathcal{D}_{r,t} := \left\|\underline{W}_l^{r,t}-\underline{W}^{r,t}\right\|_F^2$. This term is bounded as

$$\begin{aligned}
\mathcal{D}_{r,t} &\overset{(a)}{=} \mathbb{E}\left\|QPW^{r,t}-PW^{r,t}\right\|_F^2\\
&\overset{(b)}{=} \mathbb{E}\left\|(Q-P)W^{r,t}\right\|_F^2\\
&\overset{(c)}{=} \mathbb{E}\left\|(Q-P)\left(W^{r,0}-\eta\sum_{\tau=0}^{t-1}\partial\hat{\Phi}\left(W^{r,\tau}\right)\right)\right\|_F^2,
\end{aligned}$$

where (a) follows since $QPW^{r,t}=\underline{W}^{r,t}$ and $PW^{r,t}=\underline{W}_l^{r,t}$, (b) follows from $QP=Q$, and (c) follows from the update $W^{r,t}=W^{r,0}-\eta\sum_{\tau=0}^{t-1}\partial\hat{\Phi}\left(W^{r,\tau}\right)$. Using the fact that $\|a+b\|^2\leq 2\|a\|^2+2\|b\|^2$ in the above, we get

$$\begin{aligned}
\mathcal{D}_{r,t} &\leq 2\mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2+2\eta^2 t\sum_{\tau=0}^{t-1}\mathbb{E}\left\|(Q-P)\partial\hat{\Phi}\left(W^{r,\tau}\right)\right\|_F^2\\
&\leq 2\mathbb{E}\left\|(Q-P)W^{r,0}\right\|+2\eta^2 t\sum_{\tau=0}^{t-1}\lambda_2^2\mathbb{E}\|\partial\hat{\Phi}\left(W^{r,\tau}\right)\|_F^2.
\end{aligned} \tag{33}$$

The term $\mathbb{E}\left\|\partial\hat{\Phi}\left(W^{r,\tau}\right)\right\|_F^2$ in the above can be bounded as

$$\begin{aligned}
\mathbb{E}\left\|\partial\hat{\Phi}\left(W^{r,\tau}\right)\right\|_F^2 &= \mathbb{E}\sum_{k=1}^{N}\left\|\frac{1}{b}\sum_{j\in B_k^{r,t}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)\right\|_2^2\\
&\leq \mathbb{E}\sum_{k=1}^{N}\frac{1}{b}\sum_{j\in B_k^{r,t}}\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)\|_2^2\\
&\overset{(a)}{\leq} 2l_{max}\sum_{k=1}^{N}\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right],
\end{aligned}$$

where (a) follows from the smoothness assumption and the fact that $l_{max} := \max_{k,j} l_{k,j}$. Using equation 10 of Lemma 1, i.e., $\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \frac{2L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r - \underline{\boldsymbol{w}}^r\right\|_2^2 + \frac{2}{\mu_{min}}\mathbb{E}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|^2$ in the above, we get

$$\mathbb{E}\left\|\partial\hat{\Phi}\left(W^{r,\tau}\right)\right\|_F^2 \quad \leq \quad \frac{4L_{max}^2 l_{max}}{\mu_{min}}\sum_{k=1}^{N}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r - \underline{\boldsymbol{w}}^r\right\|_2^2 + \frac{4l_{max}}{\mu_{min}}\sum_{k=1}^{N}\mathbb{E}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|^2.$$

The result above can be written in the matrix form as,

$$\mathbb{E}\left\|\partial\hat{\Phi}\left(W^{r,\tau}\right)\right\|_F^2 = \frac{4L_{max}^2 l_{max}}{\mu_{min}}\mathcal{D}_{r,0} + \frac{4l_{max}}{\mu_{min}}\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r,0}\right)\right\|_F^2.$$

Substituting the above result in equation 33, we get

$$\mathcal{D}_{r,t} \quad \leq \quad 2\mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2 + 4\eta^2 L_{max}^2\lambda_2^2\gamma t^2\mathcal{D}_{r,0} + 4\eta^2\lambda_2^2\gamma t^2\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r,0}\right)\right\|_F^2, \tag{34}$$

where $\gamma := \frac{2l_{max}N}{\mu_{min}}$. Now, consider

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \quad \leq \quad & \left(1-\frac{\eta\mu}{4}\right)^T\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^r\right)\right] + \\ & \frac{2\eta L^2}{N}\sum_{\tau=0}^{T-1}\left(1-\frac{\eta\mu}{4}\right)^\tau\sum_{k=1}^{N}\mathbb{E}\left(\left\|\boldsymbol{w}_k^{r,T-1-\tau} - \underline{\boldsymbol{w}}_k^{r,T-1-\tau}\right\|^2 + \left\|\underline{\boldsymbol{w}}_k^{r,T-1-\tau} - \underline{\boldsymbol{w}}^{r,T-1-\tau}\right\|^2\right)\end{aligned}$$

where (a) follows from the fact that $\left\|\boldsymbol{w}_k^{r,T-1-\tau} - \underline{\boldsymbol{w}}_k^{r,T-1-\tau}\right\|^2 = 0$ and $\left\|\underline{\boldsymbol{w}}_k^{r,T-1-\tau} - \underline{\boldsymbol{w}}^{r,T-1-\tau}\right\|^2 = 0$ for $\tau = T-1$. Now choosing $\eta < \frac{4}{\mu}$ and uainf equation 32 and equation 34, the average loss becomes

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \quad \leq \quad & \mathbb{E}\left[\left(\left(1-\frac{\eta\mu}{4}\right)^T + \frac{64\eta^3T^3 l_{max}L^2L_{max}}{\mu_{min}}\right)\Phi\left(\underline{\boldsymbol{w}}^r\right) + \frac{4\eta TL^2}{N}\left\|(Q-P)W^{r,0}\right\|_F^2 + \right. \\ & \left.\frac{2\eta TL^2}{N}\left[\left(\frac{16l_{max}\eta^2T^2L_{max}^2}{\mu_{min}} + 4\lambda_2^2\eta^2\gamma L_{max}^2T^2\right)\mathcal{D}_{r,0} + 4\eta^2\gamma T^2\lambda_2^2\left\|\partial\Phi\left(\underline{W}^{r,0}\right)\right\|_F^2\right]\right].\end{aligned} \tag{35}$$

Using the fact that $\left(1-\frac{\eta\mu}{4}\right)^T \leq \left(1-\frac{\eta\mu}{4}\right)$, the above can be further bounded as

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \quad \leq \quad & \mathbb{E}\left[\left(1-\frac{\eta\mu}{4} + \frac{64\eta^3T^3 l_{max}L^2L_{max}}{\mu_{min}}\right)\Phi\left(\underline{\boldsymbol{w}}^r\right) + \frac{2\eta TL^2}{N}\left[2\left\|(Q-P)W^{r,0}\right\|_F^2 + \right.\right. \\ & \left.\left.\left(\frac{16l_{max}\eta^2T^2L_{max}^2}{\mu_{min}} + 4\lambda_2^2\eta^2\gamma L_{max}^2T^2\right)\mathcal{D}_{r,0} + 4\eta^2\gamma T^2\lambda_2^2\left\|\partial\Phi\left(\underline{W}^{r,0}\right)\right\|_F^2\right]\right].\end{aligned} \tag{36}$$

The term $\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r,0}\right)\right\|_F^2$ can be bounded as

$$\mathbb{E}\left\|\partial\hat{\Phi}\left(\underline{W}^{r,0}\right)\right\|_F^2 = \sum_{k=1}^{N}\mathbb{E}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|^2 \quad \overset{(a)}{\leq} \quad \sum_{k=1}^{N}2L_{max}\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right] \quad = 2L_{max}N\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^r\right)\right],$$

where (a) follows from the smoothness assumption and (b) follows from the fact that $\Phi\left(\underline{\boldsymbol{w}}^r\right) = \frac{1}{N}\sum_{k=1}^{N}\Phi_k\left(\underline{\boldsymbol{w}}^r\right)$. Using the above result in equation 36, we get

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \quad \leq \quad & \mathbb{E}\left[\left(1-\frac{\eta\mu}{4} + \frac{64\eta^3T^3 l_{max}L^2L_{max}}{\mu_{min}}\right)\Phi\left(\underline{\boldsymbol{w}}^r\right) + \frac{2\eta TL^2}{N}\left[2\left\|(Q-P)W^{r,0}\right\|_F^2 + \right.\right. \\ & \left.\left.\left(\frac{16l_{max}\eta^2T^2L_{max}^2}{\mu_{min}} + 4\lambda_2^2\eta^2\gamma L_{max}^2T^2\right)\mathcal{D}_{r,0} + 8\eta^2\lambda_2^2\gamma T^2L_{max}N\Phi\left(\underline{\boldsymbol{w}}^r\right)\right]\right] \\ \leq \quad & \mathbb{E}\left[\left(1-\frac{\eta\mu}{4} + \frac{64\eta^3T^3 l_{max}L^2L_{max}}{\mu_{min}} + 16\eta^3\gamma T^3\lambda_2^2L^2L_{max}\right)\Phi\left(\underline{\boldsymbol{w}}^r\right) + \right. \\ & \left.\frac{2\eta TL^2}{N}\left(\left[\frac{16l_{max}\eta^2T^2L_{max}^2}{\mu_{min}} + 4\lambda_2^2\eta^2\gamma L_{max}^2T^2\right]\mathcal{D}_{r,0} + 2\left\|(Q-P)W^{r,0}\right\|_F^2\right)\right].\end{aligned}$$

Choosing $\eta \leq \frac{1}{8}\left(\frac{\mu}{\frac{64T^3 l_{max} L^2 L_{max}}{\mu_{min}} + 16\gamma T^3 L^2 \lambda_2^2 L_{max}}\right)^{1/2}$ in the above result in

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] &\leq \mathbb{E}\left[\left(1-\frac{\eta\mu}{8}\right)\Phi\left(\underline{\boldsymbol{w}}^{r}\right) + \frac{2\eta^3 TL^2}{N}\left[\frac{16T^2 L_{max}^2 l_{max}}{\mu_{min}} + 4\lambda_2^2\gamma L_{max}^2 T^2\right]\mathcal{D}_{r,0}\right. \\ &+ \left.\frac{4\eta TL^2}{N}\left\|(Q-P)W^{r,0}\right\|_F^2\right]. \end{aligned} \tag{37}$$

Again choosing $\eta \leq \left(\frac{1}{\frac{16T^2 l_{max} L_{max}^2}{\mu_{min}} + 4\lambda_2^2\gamma T^2 L_{max}^2}\right)^{\frac{1}{2}}$, the above results in

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \leq \left(1-\frac{\eta\mu}{8}\right)\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r}\right)\right] + \frac{2\eta TL^2}{N}\mathcal{D}_{r,0} + \frac{4\eta TL^2}{N}\mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2.$$

It is easy to see that $\mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2 = \mathbb{E}\left\|\underline{W}_l^{r,0} - \underline{W}^{r,0}\right\|_F^2 = \mathcal{D}_{r,0}$. Using this above, gives us

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \leq \left(1-\frac{\eta\mu}{8}\right)\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r}\right)\right] + \frac{6\eta TL^2}{N}\mathcal{D}_{r,0}. \tag{38}$$

This completes the proof. □

B.3 Proof of Lemma 2

Let $\mathcal{D}_{r,0} = \mathbb{E}\left\|\underline{W}_l^{r,0} - \underline{W}^{r,0}\right\|_F^2 = \sum_{k=1}^N \mathbb{E}\left\|\underline{\boldsymbol{w}}_k^{r,0} - \underline{\boldsymbol{w}}^{r,0}\right\|^2$. Using compact notations for the updates in equation 7 and equation 8, the global drift term can be written as

$$\begin{aligned}\mathcal{D}_{r,0} &= \mathbb{E}\left\|QPW^{r,0} - PW^{r,0}\right\|_F^2 \\ &= \mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2. \end{aligned} \tag{39}$$

Recall that $Q = \frac{1}{N}\mathbf{1}\mathbf{1}^T$ is the average matrix, P is the mixing matrix and $QP = Q$. Using $\underline{W}_l^{r,0} = PW^{r-1,T}$ (see equation 7), substituting for the update in $W^{r-1,T}$ and taking the telescopic sum, we get

$$W^{r,0} = \underline{W}_l^{r,0} = P\left(W^{r-1,0} - \eta\sum_{\tau=0}^{T-1}\partial\hat{\Phi}\left(W^{r-1,\tau}\right)\right).$$

Plugging the above in equation 39, and using the generalized Cauchy's inequality, i.e., $\|a+b\|^2 \leq \left(1+\frac{1}{\psi}\right)\|a\|^2 + (1+\psi)\|b\|^2$ for any $\psi \geq 0$, the global drift term can be upper bounded as

$$\begin{aligned}\mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2 &\leq \left(1+\frac{1}{\psi}\right)\Xi + (1+\psi)\eta^2\mathbb{E}\left\|\left(Q-P^2\right)\sum_{\tau=0}^{T-1}\partial\hat{\Phi}\left(W^{r-1,\tau}\right)\right\|_F^2 \\ &\overset{(a)}{\leq} \left(1+\frac{1}{\psi}\right)\Xi + (1+\psi)\eta^2\left\|\left(Q-P^2\right)\right\|_{op}^2\mathbb{E}\left\|\sum_{\tau=0}^{T-1}\partial\hat{\Phi}\left(W^{r-1,\tau}\right)\right\|_F^2 \\ &\overset{(b)}{\leq} \left(1+\frac{1}{\psi}\right)\Xi + (1+\psi)\eta^2\lambda_2^4 T\sum_{\tau=0}^{T-1}\mathbb{E}\left\|\partial\hat{\Phi}\left(W^{r-1,\tau}\right)\right\|_F^2, \end{aligned} \tag{40}$$

where λ_2 is the second largest eigenvalue of the mixing matrix P and $\Xi := \mathbb{E}\left\|\left(Q-P^2\right)W^{r-1,0}\right\|_F^2$. In the above, (a) follows from Lemma 4 and (b) follows from Lemma 5. Next, consider bounding the following

$$\begin{aligned}
\mathbb{E}\left\|\partial\hat{\Phi}\left(W^{r-1,\tau}\right)\right\|_F^2 &= \mathbb{E}\sum_{k=1}^{N}\left\|\frac{1}{b}\sum_{j\in B_k^{r-1,\tau}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right\|_2^2 \\
&\overset{\text{Jensen's}}{\leq} \mathbb{E}\sum_{k=1}^{N}\frac{1}{b}\sum_{j\in B_k^{r-1,\tau}}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right\|_2^2 \\
&\overset{(a)}{\leq} 2l_{max}\sum_{k=1}^{N}\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right], \qquad (41)
\end{aligned}$$

where (a) follows from the smoothness assumption and $l_{max} := \max_{k,j} l_{k,j}$. Recall from Lemma 1 that

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right] \leq \frac{2L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^{r-1}-\underline{\boldsymbol{w}}^{r-1}\right\|_2^2 + \frac{2}{\mu_{min}}\mathbb{E}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^{r-1}\right)\right\|^2.$$

Substituting this in equation 41, and writing it in the matrix form, we get

$$\mathbb{E}\left\|\partial\hat{\Phi}\left(W^{r-1,\tau}\right)\right\|_F^2 = \frac{4l_{max}L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{W}_l^{r-1,0}-\underline{W}^{r-1,0}\right\|_F^2 + \frac{4l_{max}}{\mu_{min}}\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r-1,0}\right)\right\|_F^2.$$

Using the above in equation 40

$$\begin{aligned}
\mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2 &\leq \left(1+\frac{1}{\psi}\right)\mathbb{E}\left\|\left(Q-P^2\right)W^{r-1,0}\right\|_F^2 + \eta^2\lambda_2^4\alpha T^2 L_{max}^2\mathcal{D}_{r-1,0} \\
&\quad +\eta^2\lambda_2^4\beta T^2\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r-1,0}\right)\right\|_F^2, \qquad (42)
\end{aligned}$$

where $\beta := \frac{4l_{max}(1+\psi)}{\mu_{min}}$. First, let us consider bounding a part of the first term above, i.e., $\mathbb{E}\left\|\left(Q-P^2\right)W^{r-1,0}\right\|_F^2$. Using the fact that $QP=Q$ and $Q^2=Q$, it follows that $P^2-Q=(Q-P)^2$. Using this in equation 42, we get

$$\begin{aligned}
\mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2 &\leq \left(1+\frac{1}{\psi}\right)\mathbb{E}\left\|(Q-P)^2W^{r-1,0}\right\|_F^2 + \eta^2\lambda_2^4\beta T^2 L_{max}^2\mathcal{D}_{r-1,0} \\
&\quad +\eta^2\lambda_2^4\beta T^2\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r-1,0}\right)\right\|_F^2.
\end{aligned}$$

Applying the results of Lemma 4 and 5, the first term in the above can further be bounded as,

$$\begin{aligned}
\mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2 &\leq \left(1+\frac{1}{\psi}\right)\|(Q-P)\|^2\,\mathbb{E}\left\|(Q-P)W^{r-1,0}\right\|_F^2 + \eta^2\lambda_2^4\beta T^2 L_{max}^2\mathcal{D}_{r-1,0} \\
&\quad +\eta^2\lambda_2^4\beta T^2\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r-1,0}\right)\right\|_F^2 \\
&\overset{(a)}{\leq} \left(1+\frac{1}{\psi}\right)\lambda_2^2\mathcal{D}_{r-1,0} + \eta^2\lambda_2^4\beta T^2 L_{max}^2\mathcal{D}_{r-1,0} + \eta^2\lambda_2^4\beta T^2\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r-1,0}\right)\right\|_F^2, \qquad (43)
\end{aligned}$$

where (a) follows by substituting the results from Lemma 5. The term $\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r-1,0}\right)\right\|_F^2$ in the above is bounded as follows

$$\begin{aligned}
\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r-1,0}\right)\right\|_F^2 &= \mathbb{E}\sum_{k=1}^{N}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right\|_2^2 \\
&\overset{(a)}{\leq} 2L_{max}N\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right],
\end{aligned}$$

where (a) follows from smoothness assumption and using the fact that $\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right) = \frac{1}{N}\sum_{k=1}^{N}\Phi_k\left(\underline{\boldsymbol{w}}^{r-1,0}\right)$, and $L_{max} = \max_k L_k$. Using the above result in equation 43, we get

$$\mathcal{D}_{r,0} \quad \leq \quad \left(\left(1+\frac{1}{\psi}\right)\lambda_2^2 + \eta^2\lambda_2^4\beta T^2 L_{max}^2\right)\mathcal{D}_{r-1,0} + 2\eta^2\lambda_2^4\beta T^2 L_{max} N\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right].$$

This completes the proof. □

B.4 Proof of Lemma 3 (Completing the proof of Theorem 1)

Let us recall the equations for $\mathcal{D}_{r+1,0}$ and $\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)$ from Lemma 1 and Lemma 2

$$\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right) \quad \leq \quad \alpha\Phi\left(\underline{\boldsymbol{w}}^{r}\right) + \rho\mathcal{D}_{r,0}, \tag{44}$$

$$\mathcal{D}_{r+1,0} \quad \leq \quad \nu\mathcal{D}_{r,0} + \chi\Phi\left(\underline{\boldsymbol{w}}^{r}\right), \tag{45}$$

where $\alpha := \left(1-\frac{\eta\mu}{8}\right), \rho := \frac{6\eta L^2 T}{N}, \nu := \left(1+\frac{1}{\psi}\right)\lambda_2^2 + \eta^2\lambda_2^4\beta T^2 L_{max}^2$ and $\chi := 2\eta^2\lambda_2^4\beta T^2 L_{max} N$. To ensure $\nu < 1$, we choose $\psi = \frac{2\lambda_2^2}{1-\lambda_2^2}$ and any $\eta \leq \sqrt{\frac{1-\lambda_2^2}{4\lambda_2^4\beta T^2 L_{max}^2}}$.Further, to ensure $\chi < 1$, we choose $\eta \leq \sqrt{\frac{1}{2\lambda_2^4\beta T^2 L_{max} N}}$. Now consider the following Lyapunov function for some constant $\theta > 0$

$$\begin{aligned}\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right) + \theta\mathcal{D}_{r+1,0} \quad &\leq \quad \alpha\Phi\left(\underline{\boldsymbol{w}}^{r}\right) + \rho\mathcal{D}_{r,0} + \theta\left(\nu\mathcal{D}_{r,0} + \chi\Phi\left(\underline{\boldsymbol{w}}^{r}\right)\right) \\ &\overset{(a)}{\leq} \quad \left(\alpha+\theta\chi\right)\Phi\left(\underline{\boldsymbol{w}}^{r}\right) + \left(\rho+\theta\nu\right)\mathcal{D}_{r,0},\end{aligned} \tag{46}$$

where (a) follows from equation 44 and equation 45. To show linear convergence we want the coefficients of the first and second terms in equation 46 to satisfy the following inequalities

$$\alpha + \theta\chi \leq \left(1-\frac{\eta\mu}{16}\right) \text{ and } (\rho+\theta\nu) \leq \theta\left(1-\frac{\eta\mu}{16}\right). \tag{47}$$

Now, consider the first inequality above. Substituting for α and χ and choosing

$$\eta \leq \frac{\mu}{32\theta\lambda_2^4\beta T^2 L_{max} N},$$

ensures that the first inequality in equation 47 is satisfied. Next, substituting the values for ρ and ν in the second inequality in equation 47, and simplifying results in

$$\frac{6\eta L^2 T}{N} + \theta\eta^2\lambda_2^4\beta T^2 L_{max}^2 + \frac{\theta\eta\mu}{16} \quad \leq \quad \theta\left(1-\lambda_2^2-\frac{\lambda_2^2}{\psi}\right),$$

where the above quantity is non-negative by choosing $\psi > \frac{\lambda_2^2}{1-\lambda_2^2}$. Now, picking $\eta \leq \frac{1}{\theta\lambda_2^4\beta T^2 L_{max}^2}$ leads to

$$\eta\left(1+\frac{\theta\mu}{16}+\frac{6L^2T}{N}\right) \quad \leq \quad \theta\left(1-\lambda_2^2-\frac{\lambda_2^2}{\psi}\right).$$

Choosing $\eta \leq \frac{\theta\left(1-\lambda_2^2-\frac{\lambda_2^2}{\psi}\right)}{\left(1+\frac{\eta\mu}{16}+\frac{6L^2T}{N}\right)}$ ensures that $(\rho+\theta\nu) \leq \theta\left(1-\frac{\eta\mu}{16}\right)$. Finally, choosing

$$\begin{aligned}\eta \leq \min\Bigg\{&\frac{4}{\mu}, \frac{2}{\mu_{min}}, \frac{\mu}{4\zeta_1}, \frac{L^2}{2\zeta_2}, \frac{1}{8}\left(\frac{\mu}{\zeta_3 T^3}\right)^{\frac{1}{3}}, \left(\frac{1}{\zeta_4 T^2}\right)^{\frac{1}{2}}, \frac{\mu_{min}}{\zeta_5}, \\ &\sqrt{\frac{1-\lambda_2^2}{\zeta_6 T^2}}, \sqrt{\frac{1}{\zeta_7 T^2}}, \frac{\mu}{\zeta_8 T^2}, \frac{1}{\zeta_9 T^2}, \frac{\theta(1-\lambda_2^2-\frac{\lambda_2^2}{\psi})}{\left(1+\frac{\theta\mu}{16}+\frac{6L^2T}{N}\right)}\Bigg\},\end{aligned} \tag{48}$$

Substituting the conditions in equation 47 in equation 46, we get

$$\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right) + \theta\mathcal{D}_{r+1,0} \quad \leq \quad \left(1-\frac{\eta\mu}{16}\right)\left(\Phi\left(\underline{\boldsymbol{w}}^{r}\right) + \theta\mathcal{D}_{r,0}\right).$$

for any constant $\theta > 0$. Here, $\mu_{min} := \min_{k\in[N]}\{\mu_k\}$, $l_{max} := \max_{k,j} l_{k,j}$ and $L_{max} := \max_k L_k$, $\psi = \frac{2\lambda_2^2}{1-\lambda_2^2}$, $\gamma := \frac{2l_{max}N}{\mu_{min}}$, $\beta := \frac{4l_{max}(1+\psi)}{\mu_{min}}$, and $\mathcal{D}_{r,0} := \sum_{k=1}^{N} \mathbb{E}\left\|\underline{\boldsymbol{w}}_k^{r,0} - \underline{\boldsymbol{w}}^{r,0}\right\|^2$. Moreover, in the above $\zeta_1 := 4\left(\frac{2Ll_{max}}{bN} + \frac{2LL_{max}}{N} + 2LL_{max}\right)$, $\zeta_2 := 2\left(\frac{Ll_{max}^2}{bN} + \frac{LL_{max}^2}{N} + LL_{max}^2\right)$, $\zeta_3 := \frac{64l_{max}LL_{max}}{\mu_{min}} + 16\gamma L\lambda_2^2 L_{max}$, $\zeta_4 := \frac{16l_{max}L_{max}^2}{\mu_{min}} + 4\lambda_2^2\gamma L^2$, $\zeta_5 := 2\left[\frac{l_{max}L_{max}}{b} + \frac{L_{max}^2 b(b-1)}{b^2}\right]$, $\zeta_6 := 4\lambda_2^4\beta L_{max}^2$, $\zeta_7 := 2\lambda_2^4\beta L_{max}N$, $\zeta_8 := 32\theta\lambda_2^4\beta L_{max}N$ and $\zeta_9 := \theta\lambda_2^4\beta L_{max}^2$.

□

C Differences between strongly convex and our setting

In the following, we provide a very simple 1-D examples for strongly convex ($f_1(x) = \log(x - 0.5 + \sqrt{1+(x-0.5)^2}) + (x-0.5)^2 + 0.25 +$ constant and $f_2(x) = 2x^2$), and non-convex settings ($f_1(x) = x^2 + 2\sin^2(x)$ and $f_2(x) = 0 \times \mathbf{1}\{f_1(x) \leq 4\} + f_1(x)\mathbf{1}\{f_1(x) > 4\}$), as shown in Fig. 8. Note that in the strongly convex setting, both clients share the unique minima $x^* = 0$ due to the interpolation assumption. In this case, both clients do not need to communicate since each client can run local rounds to reach the global minima that minimize the average, and hence making decentralized or collaborative learning vacuous! On the other hand, for the PL setting, multiple local rounds lead to different optimal points. For example, running multiple rounds of GD results in client 1 reaching $x_1^* = 0$ while client 2 reaches ≈ -1.3 or $\approx +1.3$ depending on the initialization. However, the optimal point $x^* = 0$. This simple scenario suggests challenges while proving the results. For example, in the strongly convex setting ((Koloskova et al., 2020)), one can start with the difference between global optimum and the local/global update while we cannot use this to prove our results.

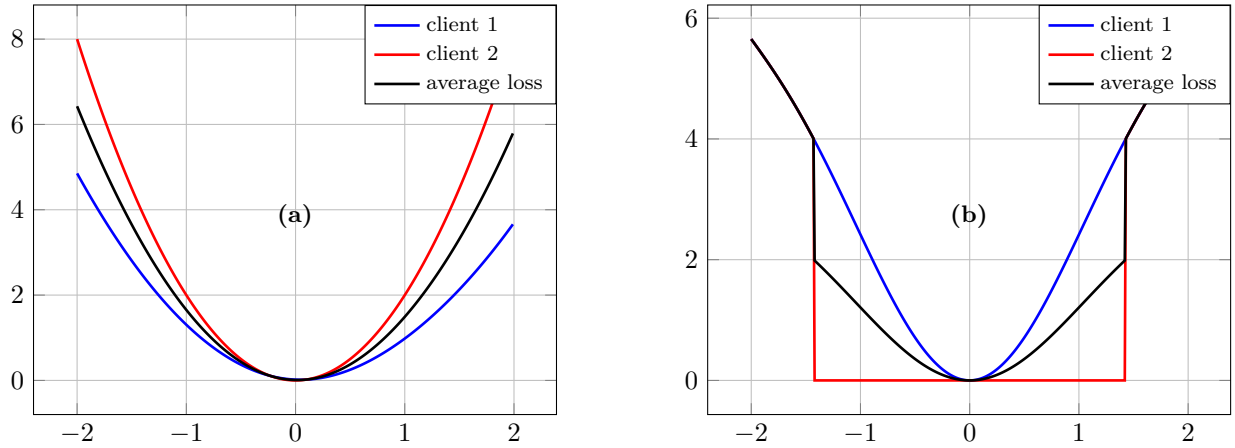

Figure 8: Strongly convex losses in the overparameterized regime (see (a)) and Losses satisfying PL inequality (see (b)).